Juarez Angelo Piazza Sacenti

ADAPTAÇÃO DE HIERARQUIAS DE DADOS CONECTADOS PARA ANÁLISE DE INFORMAÇÃO

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Ciência da Computação.
Orientador: Prof. Renato Fileto, Dr.

Florianópolis

2016

Ficha de identificação da obra elaborada pelo autor através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

A ficha de identificação é elaborada pelo próprio autor

INSERIDA APENAS NA VERSÃO FINAL

Maiores informações em:
http://portalbu.ufsc.br/ficha

Juarez Angelo Piazza Sacenti

ADAPTAÇÃO DE HIERARQUIAS DE DADOS CONECTADOS PARA ANÁLISE DE INFORMAÇÃO

Esta dissertação foi julgada adequada para obtenção do título de mestre e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

Florianópolis, 23 de fevereiro 2016.

Prof. Carina Friedrich Dorneles, Dr[a].
Coordenadora do Programa

Banca Examinadora:

Prof. Renato Fileto, Dr.
Universidade Federal de Santa Catarina
Orientador

Prof. José Antonio Fernandes de Macedo, Dr.
Universidade Federal do Ceará

Prof. Denilson Sell, Dr.
Universidade Federal de Santa Catarina

Prof. Roberto Willrich, Dr.
Universidade Federal de Santa Catarina

Prof. Mario Antonio Ribeiro Dantas, Dr.
Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus pais, familiares e todos aqueles que incentivaram-me a continuar estudando.

AGRADECIMENTOS

Grato a paciente orientação de Renato Fileto e toda sua dedicação proporcionar as melhores oportunidades de aprendizado e crescimento para formar futuros pesquisadores.

Agradeço aos meus colegas do LISA (Laboratório para Integração de Sistema e Aplicações Avançadas) que criaram um ambiente de amizade, integração, e conhecimento. Entre eles, aos meus amigos: André Salvaro, Areli Santos, Cleto May, Douglas Klein, Felipe Pinto, Felipe Born, Filipe Silva, Geomar Schreiner, Jean Gonçalves, Lucas de Alencar, Ramon Hugo de Souza e Ricardo Nabo, pelas críticas construtivas, auxílio e companheirismo. Em especial aos meus amigos: Douglas Klein, pela coleta e mineração textual em *tweets* e a Cleto May, pela associação semântica de *tweets* com dados de movimento, ambas tarefas essenciais para este trabalho.

Agradecimentos a Alessandra Rafaettà, Alessandro Roncato, Fabio Salvini pela orientação, contribuição acadêmica e acolhimento caloroso por suas famílias em terras extrangeiras. Obrigado ao companheirismo de Chiara Gaetani e Edson J. M. Fumagalli durante a estadia na *bella Italia*.

Grato a toda minha família, em especial: a minha mãe, Mirian Célia Piazza Sacenti, por todo amor, caridade, entrega, resignação e carinho; ao meu pai, Juarez Angelo Sacenti, por toda dedicação e compreensão pelas minhas escolhas; a minhas tia Odete Maria de Olveira por todo carinho, orientação e instrução, muito muito obrigado; a minha tia Doroti Sacenti pelo carinho e conforto; a meu tio Mario Sérgio Piazza pela fidelidade nas caminhadas de fim de tarde durante os últimos meses do mestrado.

Por fim, agradeço a banca examinadora deste trabalho por suas críticas, considerações e indicações de trabalho futuros.

Corcovado, the hill in Rio where stands the statue of Christ the Redeemer, can be categorized (typed) as instance of *Mountain* and *Touristic Place* at the same time.
(FILETO et al., 2014)

RESUMO

Diversas abordagens têm sido propostas para o enriquecimento semântico de dados sobre movimento, incluindo propostas para a sua anotação com dados abertos conectados (LOD). Contudo, ainda há uma carência de soluções para modelagem dimensional de dados semanticamente anotados, visando sua análise em *data warehouses*. Este trabalho de pesquisa propõe um método para a geração automatizada de dimensões de análise de dados a partir da adaptação de hierarquias de recursos (sobre instâncias e conceitos) de LOD usados para anotar semanticamente tais dados. Este método extrai hierarquias de recursos de coleções de LOD por meio da exploração de relações de ordenamento parcial (como *part of* e *is a*) e adapta tais hierarquias, reduzindo o número de recursos de acordo com o número de vezes que um recurso aparece em anotações semânticas de uma dada coleção de dados. Dimensões assim produzidas são potencialmente menores que a hierarquia usada para gerá-las, pois escondem recursos com baixa frequência de uso em anotações. Isso tem potencial para propiciar ganhos de eficiência e facilidade de uso em *data warehouses*, entre outros benefícios. Resultados de experimentos com a adaptação de hierarquias de recursos para a análise de *tweets* anotados com LOD sugerem a viabilidade do método. Os resultados mostram considerável redução no número de recursos de hierarquias adaptadas a medida que se aumenta o limiar de frequência de uso de recursos em anotações semânticas.

Palavras-chave: Dados sobre movimento. *Data Warehouse*. Dimensões de análise. *Web* semântica. Dados abertos conectados (LOD). Mídias Sociais. *Tweets* geo-localizados.

ABSTRACT

Several approaches have been proposed to semantic enrich data about movement, including proposals to annotate it with linked open data (LOD). However, there is still a lack of solutions for multidimensional modelling semantic annotated data, in order to analyse it in data warehouses. This research work proposes a method for automated generation of data analysis dimensions from the adaptation of hierarchies of resources (about instances and concepts) of LOD used to annotate semantically such data. This method extract hierarchies of resources from LOD collections through exploration of partial ordering relations (like *part of* and *is a*) and adapt such hierarchies, reducing the number of resources according to the number of times a resource appears in semantic annotations of a particular dataset. Thus produced dimensions are potentially lower than the hierarchy used to generate them, because they hide resources having low frequency of use in annotations. This has the potential to provide efficiency gains and ease of use in data warehouses, among other benefits. Experiments results in adaptation of hierarchies of resources for the analysis of tweets annotated with LOD suggest the feasibility of the method. The results show considerable reduction of the number of adapted hierarchies' resources as it increases the threshold of frequency of use of resources in semantic annotations.

Keywords: Movement data. Data warehouses. Analysis dimensions. Semantic Web. Linked open data. Social media. Geo-located tweets.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LISTA DE SÍMBOLOS

SUMÁRIO

1 INTRODUÇÃO

As tecnologias de posicionamento e comunicação (*e.g.*, GPS, GSM) possibilitam o acúmulo de grandes volumes de dados sobre movimento (*movement data*), tais como trajetórias de objetos móveis (*moving objects* - MOs) (PARENT et al., 2013; PELEKIS; THEODORIDIS, 2014) ou trilhas (*trails*), *i.e.* sequências de interações geo-localizadas de um usuário com um sistema de informação (*e.g.*, *posts* de usuários em mídias sociais, tais como Facebook[1], Twitter[2] ou Instagram[3]; ou sistemas que produzam *Web* ou *mobile logs* geo-localizados) (FILETO et al., 2015). Novos métodos têm sido propostos para enriquecer semanticamente dados sobre movimento, *e.g.* através de sua anotação com recursos sobre instâncias (objetos) e conceitos (classes de objetos) descritos em bases de conhecimento (*Knowledge Bases* - KBs). Anotações sêmanticas com recursos de KBs associam informação aos dados sobre movimento, tais como: lugares de interesse (*Place of Interest* - PoI) visitados, meios de transporte empregados, atividades realizadas, objetivos de paradas (*stops*) e deslocamentos (*moves*) (YAN et al., 2013; FILETO et al., 2013; BOGORNY et al., 2014; MAY; FILETO, 2014).

Conjuntos de dados sobre movimento semanticamente anotados (*Semantically annotated Movement Dataset* - SMoD) oportunizam a construção de *Movement Data Warehouses* (MDWs), *i.e. Data Warehouses* (DW) para a análise de dados sobre movimento, que explorem a informação associada por anotações semânticas em dimensões de análise. Dimensões de análise são hierarquias para organização e análise dos fatos de um DW em diversos níveis de abstração (CABIBBO; TORLONE, 1998). Diversos campos de aplicação beneficiam-se da análise de dados sobre movimento, tais como a gestão de tráfego, segurança urbana, *marketing* geográfico e estudos de comportamento social.

Por exemplo, considere a análise de um trecho da trilha de usuário do Twitter, ilustrada na Figura 1. Os rótulos associados a certas posições da trilha representam anotações semânticas que explicitam o tipo e o nome de PoI visitados (*e.g.*, *BusStation::TICEN*, *Memorial::Monumento ao Soldado*, *Supermarket::Angeloni Agronômica*). As dimensões espaciais de análise *Spatial Object Dim* e *Spatial Concept Dim* são hierarquias de recursos que exploram relações de ordenamento parcial (respectivamente, *part of* e *is a*) de recursos de en-

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] https://instagram.com/

riquecimento semântico referenciados em anotações semânticas. Por exemplo, o lugar visitado *Supermarket::Angeloni Agronômica* possui a cadeia de relações *part of* "*District::Agrônomica - City::Florianópolis - State::Santa Catarina - Country::Brazil*". Os conceitos *PoI / Address*, *District*, *City*, *State* e *Country* compõem os níveis de hierarquia da dimensão de análise *Spatial Object Dim*. Por outro lado, a dimensão *Spatial Concept Dim* não apresenta conceitos bem definidos para classificar seus níveis e por isso é representada em forma de árvore.

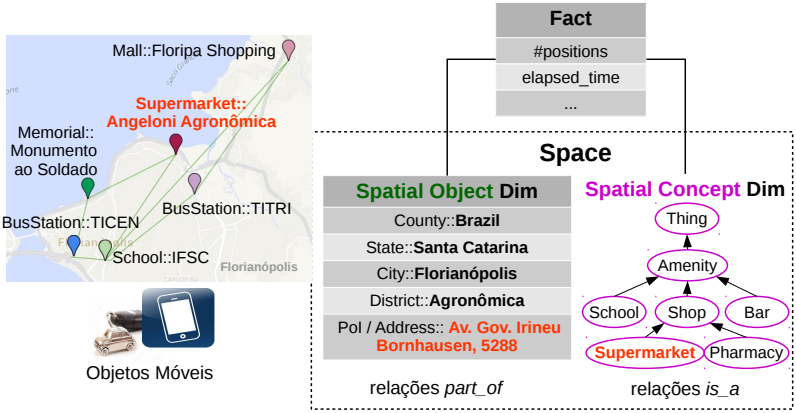


Figura 1 – Exemplo da análise de trilha de usuário

A análise de *tweets* geo-localizados, como no exemplo anterior, possibilita responder questões como (HONG et al., 2012):

- Como a informação é criada e compartilhada em diferentes lugares geográficos? Como o conteúdo textual varia de acordo com o espaço geográfico?

- Quais são as características espaciais e linguísticas das pessoas? Como a linguagem varia de acordo com a região e tipo de lugar?

- Quais são os padrões de movimentação dos usuários ao utilizar o serviço?

Além destas, dimensões espaciais de análise permitem consultas como: **"Qual o tempo decorrido (*elapsed time*) de MOs em locais da classe *Shop*?"**. Nesta consulta, a resposta é obtida pela soma do tempo de visita de cada MO em locais da classe *Shop*,

tal como o supermercado *Supermarket::Angeloni Agronômica*[4]. *Supermarket::Angeloni Agronômica* é instância de *Supermarket* e, consequentemente, é também instância da classe *Shop*, de acordo com *Spatial Concept Dim.*

1.1 DEFINIÇÃO DO PROBLEMA E DELINEAMENTO DA PROPOSTA

A construção do MDW ilustrado na Figura 1 exige, além da acomodação de SMoDs no modelo dimensional (FILETO et al., 2014), métodos apropriados para explorar a informação disponível sobre os recursos de KBs associados pelas anotações semânticas. E além disso, ainda há uma carência de soluções para a geração de dimensões de análise para dados anotados com recursos de dados conectados abertos (*Linked Open Data* - LOD), assim como trabalhos que descrevam a construção de *data warehouses* como o MDW exemplificado acima. Dado o problema da geração de dimensões de análise de dados semanticamente anotados com LOD, este trabalho propõe as seguintes hipóteses:

Hipótese 1. Dados sobre movimento anotados com recursos de LOD podem ser analisados em *data warehouses*, ao utilizar hierarquias de recursos construídas a partir de relações de ordem parcial (*e.g.*, *part of*, *is a*) descritas em coleções de LOD.

Hipótese 2. Hierarquias de recursos podem ser adaptadas com o intuito de reduzir o número de recursos da hierarquia, considerando o número de dados alvos de anotações que cada recurso da hierarquia anotou (*i.e.* frequência de uso do recurso em anotações semânticas).

Hipótese 3. Hierarquias de conceitos baseadas em relações *is a* (*subsumption*), quando utilizadas como dimensões de análise, proporcionam novos meios de analisar conjuntos de dados semanticamente anotados.

Este trabalho propõe um método para a geração automatizada de dimensões de análise de dados a partir da adaptação de hierarquias de recursos (instâncias e conceitos) conectados a recursos usados em anotações semânticas. Este método explora hierarquias derivadas de propriedades existentes em coleções de LOD e reduz o número de recursos das hierarquias de acordo com o número de anotações semânticas a que cada recurso está (direta ou indiretamente) relacionado. O método

[4] representado pelo recurso `http://linkedgeodata.org/triplify/way73321157`

é validado utilizando uma base de *tweets* semanticamente anotados com recursos de KBs.

Dimensões assim produzidas são potencialmente menores que a hierarquia usada para gerá-las, pois ocultam recursos pouco usados em anotações. A fase de adaptação de hierarquias proporciona ganhos em eficiência e facilidade no uso das dimensões de análise em *data warehouses*, entre outros benefícios, tal como a disponibilização de dimensões de análise baseadas em hierarquias de conceitos.

O método proposto suporta a análise de qualquer coleção de dados semanticamente anotados, porém os estudos de caso limitam-se a dados sobre movimento semanticamente anotados devido ao interesse atual do grupo de pesquisa no qual este trabalho foi realizado. O desenvolvimento de um MDW como do exemplo proposto foi um objetivo não alcançado deste trabalho, devido a limitações de escopo e tempo. Experimentos com outros tipos de dados e o desenvolvimento do MDW são sugeridos como futuros trabalhos.

1.2 OBJETIVOS

O objetivo geral deste trabalho de pesquisa é contribuir para a modelagem dimensional de dados de movimento anotados com recursos (instâncias e conceitos) de KBs. Especificamente, esta dissertação propõe um método para a geração automatizada de dimensões de análise a partir de conjuntos de dados semanticamente anotados e por meio da adaptação de hierarquias de recursos.

Os objetivos específicos desta dissertação são:

1. Estabelecer estratégia para gerar hierarquias de recursos (instâncias e conceitos) a partir de recursos referenciados por conjuntos de dados semanticamente anotados.

2. Estabelecer algoritmo para adaptação de hierarquias de recursos considerando o número de dados alvos de anotações que cada recurso da hierarquia anotou.

3. Analisar os efeitos da adaptação de hierarquias de recursos.

1.3 MATERIAL E MÉTODOS

O método empregado nesta dissertação compreende os seguintes passos:

1. Levantamento bibliográfico nas áreas de análise de dados sobre movimento em DW, uso de tecnologias da *Web* semântica para construção de DW, e análise de informação de mídias sociais em DW;

2. Desenvolver um método de geração de dimensões de análise de SMoDs a partir de hierarquias de recursos;

3. Definir e implementar algoritmos de extração e adaptação de hierarquias de recursos a partir de SMoDs;

4. Obter SMoDs para realização de experimentos: utilizar-se-á dados gerados por trabalhos anteriores (MAY; FILETO, 2014);

5. Realizar experimentos para analisar os efeitos da adaptação de hierarquias de recursos extraídas a partir de diferentes SMoDs;

6. Escrita de um artigo científico relacionado ao trabalho proposto e publicação do mesmo em evento com Qualis-CC CAPES, com estrato superior ou equivalente a B3;

7. Escrita da dissertação.

1.4 ESTRUTURA DO TRABALHO

O restante deste trabalho é estruturado em 4 capítulos. O capítulo 2 define fundamentos teóricos a respeito de dados sobre movimento, anotações, *Web* semântica, LOD e *data warehouses*. O capítulo 3 apresenta as definições básicas necessárias para o entendimento da proposta e o método proposto para a extração e adaptação de hierarquias de recursos. O capítulo 4 ilustra a utilização do protótipo ferramental *SeMovDim* para a adaptação de quatro SMoDs de *tweets* anotados com recursos sobre PoI visitados pelo usuário autor do *tweet*. O capítulo 5 apresenta e compara trabalhos relacionados. O capítulo 6 apresenta as conclusões obtidas durante a pesquisa e enumera trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo define a representação de dados sobre movimento em diferentes níveis de abstração e como estes podem ser semanticamente anotados. Ele também descreve os elementos básicos de um *data warehouse*.

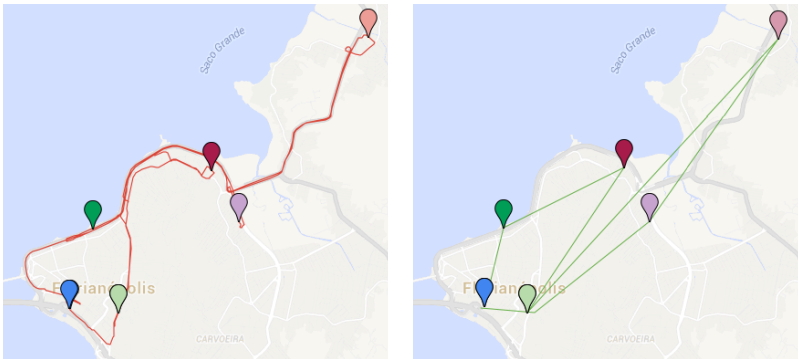
2.1 DADOS SOBRE MOVIMENTO

O termo dado sobre movimento se refere a todos os dados sobre o movimento de objetos móveis (MOs). Um conjunto de dados brutos sobre movimento (*Movement Dataset* - MoD) é um conjunto de segmentos de dados sobre movimento como o formalmente descrito pela Definição 1.

Definição 1. Um **segmento de dados brutos sobre movimento** (*Movement Data Segment* - MDS) é uma sequência temporalmente ordenada $mds = \langle p_1, \ldots, p_n \rangle$ ($n \in \mathbb{N}^+$) de posições espaço-temporais de um objeto móvel MO, onde cada posição é uma tupla $p_i = (x_i, y_i, t_i)$ com $1 \leq i \leq n$, onde (x_i, y_i) são coordenadas geográficas e t_i é um *timestamp* indicando o instante quando MO ocupava tais coordenadas.

Por exemplo, a Figura 2 ilustra dois MDSs. A Figura 2(a) apresenta uma trajetória bruta (*raw trajectory*) coletada por meio de um aplicativo baseado em tecnologias de posicionamento (*e.g.*, GPS). A Figura 2(b) mostra uma trilha (*trail*), *i.e.* sequência de interações geolocalizadas de um usuário de mídia social. Na Figura 2(a) os balões representam paradas (stops) enquanto que na Figura 2(b) eles representam posições donde foram enviadas postagens. Note que trajetórias possuem alta precisão espaçotemporal devido a taxa de amostragem regular e curta (*e.g.*, a cada poucos segundos), enquanto trilhas possuem taxa de amostragem variada, devido a natureza assíncrona de postagens dos usuários (*e.g.*, em mídias sociais). Por outro lado, postagens em mídias sociais usualmente vêm acompanhadas de diversas informações adicionais, tais como perfis de usuário e conteúdos textuais, que apesar de imprecisos podem indicar o estado e situação do MO em dada coordenada.

Ainda, um MDS pode representar subtrajetórias e trajetórias estruturadas por episódios como *stops* e *moves* (YAN et al., 2013). Entretanto, um MDS não suporta a descrição de características comuns



(a) Trajetória bruta (b) Trilha social

Figura 2 – Segmentos de dado sobre movimento (MDS)

da sequência de posições representada por um segmento (características do MO, da localidade e/ou do momento, como atividades planejadas e realizadas, objetivos, meios de transporte empregados, condições ambientais relevantes, PoI visitados, participação em eventos), por vezes importantes para a interpretação do movimento.

Similarmente a dados meramente espaciais, dados espaçotemporais sobre movimento necessitam de dois componentes para representar o movimento de um objeto móvel (RIGAUX; SCHOLL; VOISARD, 2000):

- Atributo espaço-temporal: descreve a localização, forma, orientação e tamanho do movimento de um objeto móvel no espaço, de duas ou três dimensões, e no tempo, dentro da duração do movimento.
- Atributo temático ou descritivo: descreve características do movimento por meio de atributos alfanuméricos e/ou referências a recursos semânticos, *i.e.* conceitos (classes) e objetos (instâncias de classes) definidos em uma base de conhecimento (KB).

Neste trabalho, o atributo temático enriquece semânticamente o MDS e é representado por meio de anotações.

2.2 ANOTAÇÕES

A anotação digital, também chamada de rótulo (*label*) ou etiqueta (*tag*), é uma associação de uma informação sobreposta (*supe-*

rimposed information) (DELCAMBRE; MAIER, 1999) a uma informação base (alvo da anotação), que ajuda a, por exemplo, explicar, avaliar, corrigir ou refutar o alvo. A abordagem de anotação facilita a administração das diferentes características, perspectivas e interpretações do dado sobre movimento. Um MDS representando parte (subsegmento) ou todo o movimento de um objeto móvel pode ser anotado com atributos temáticos, pelo uso de associações formalmente descritas pela Definição 2.

Definição 2. Uma **associação** é uma tupla $a = (mds, rel, at)$, onde mds é um segmento de dado sobre movimento, rel é uma relação semântica e at é um atributo temático (*e.g.*, um literal ou uma referência a um recursos descrito em base de conhecimento). Uma associação descreve uma característica comum a todas as posições espaço-temporais do MO descritas por mds (*e.g.*, local visitado em uma parada – *stop*, meio de transporte usado em um movimento – *move*).

Por exemplo, a Figura 3 apresenta o MDS $mds' = \langle p_1 ... p_8 \rangle$, uma sequência de *tweets* do usuário *@somebody*. A associação $a' = (mds'', \textit{associated\_text}, \textit{"I'm at Angeloni"})$ define a anotação de mds'' com o conteúdo textual *"I'm at Angeloni"*. Deste modo, $mds'' = \langle p_3 \rangle$ e $mds'' \subset mds'$, *i.e.* mds'' é subsegmento de mds'. Outro exemplo, a associação $a'' = (mds'', \textit{labeled}, \textit{"Angeloni"})$ define a anotação de mds'' com o conteúdo textual *"Angeloni"*, utilizado como rótulo de mds'' no mapa.

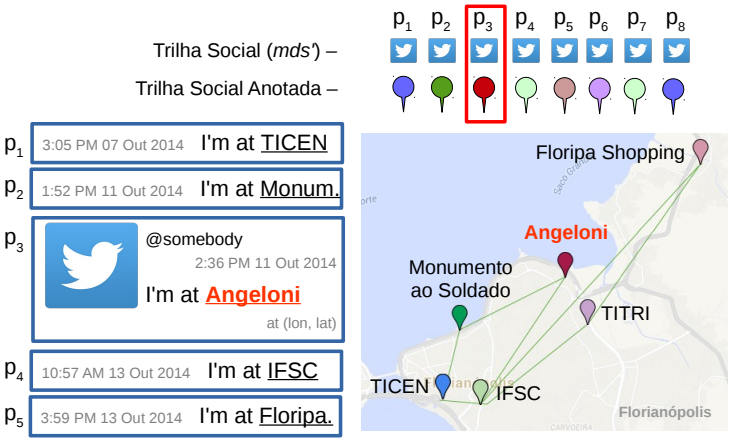


Figura 3 – Trilha anotada com PoI visitados

Anotações podem ser classificadas de acordo com a forma de armazenamento da informação sobreposta, como intrusivas ou não intrusivas (OREN et al., 2006). Anotações intrusivas são acopladas a informação base, enquanto as não intrusivas são armazenadas separadamente e referenciam a informação base com auxílio de identificadores e endereçamentos. Outra classificação de anotações considera o modo de representação da informação sobreposta, podendo ser livre ou semântica (BERNERS-LEE; HENDLER; LASSILA, 2001).

A anotação livre associa a informação base com um texto ou outra informação escrita livremente. Por não apresentar nenhum tipo de estruturação de seu conteúdo, a anotação livre é mais versátil e fácil de ser coletada. Entretanto, a liberdade da anotação livre acarreta problemas como: sinônimos, ambiguidades e erros ortográficos, tornando complexa a interpretação e utilização em aplicações.

As anotações semânticas, por outro lado, associam a informação base a descrições com semântica bem definida, referenciando a informação sobreposta com o auxílio de recursos de bases de conhecimento, ontologias (GUARINO, 1998) ou coleções dados conectados (*linked data*). A representação da anotação semântica, quando intrusiva, é também chamada de *atributo semântico*.

Diversos métodos tem sido propostos para descrever com maior exatidão o significado de anotações textuais livres (*e.g.*, a' e a''), tais como os métodos propostos por (MAY; FILETO, 2014). Usando tais métodos, é possível obter anotações semânticas como $a''' = (mds'', visits, lgd{:}way73321157)$, sendo *lgd:way73321157*[1] o recurso sobre *Supermarket::Angeloni Agronômica* descrito no LOD LinkedGeoData[2].

2.3 WEB SEMÂNTICA E LOD

Web Semântica é uma subárea de pesquisa de banco de dados que objetiva extender o papel dos computadores no suporte de diversas atividades humanas, por meio da descrição, composição e recuperação de dados e serviços que suportam diversas aplicações. A *Web* semântica fundamenta-se na utilização de tecnologias como: anotações semânticas, ontologias, bases de conhecimento, dados abertos conectados.

Uma ontologia é uma especificação explícita de uma conceituализação (GRUBER, 1995), em um ou mais domínios de conhecimento.

[1] lgd é prefixo para http://linkedgeodata.org/triplify/
[2] http://linkedgeodata.org/

As primitivas da representação de conhecimento em ontologias são conceitos (classes), objetos (instâncias) e propriedades (*i.e.* relações entre conceitos, instâncias e valores alfanuméricos).

Por exemplo, a Figura 4 apresenta uma visão ontológica (*i.e.* extrato de ontologia) composta de trechos das KBs DBpedia, LinkedGeoData (LGD) e GeoNames. Elipses verdes representam instâncias, elipses roxas representam conceitos, retângulos representam literais (*i.e.* valores alfanuméricos) e as arestas indicam propriedades. A especificação de uma ontologia se divide em dois níveis: o intencional e o nível extensional. No nível intencional são definidos os conceitos do universo de discurso, as relações entre conceitos (*e.g.*, hierarquias de classes (*subsumption*)) e suas propriedades (*e.g.*, tipos de comida servidos em restaurantes). O nível extensional descreve instâncias de acordo com o que é previsto pelo nível intencional (*e.g.*, um restaurante serve certos tipos de comida).

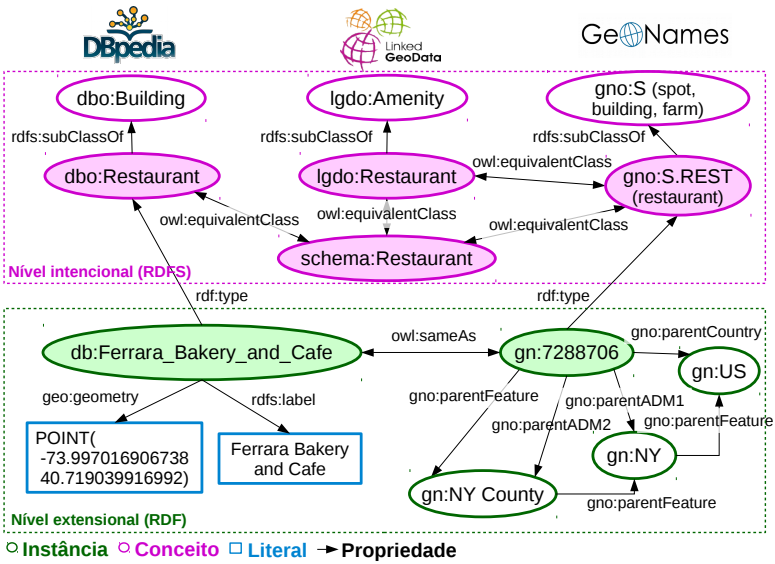


Figura 4 – Instâncias, conceitos e propriedades das coleções de dados abertos conectados DBpedia, LGD e GeoNames

Na Figura 4, pode-se observar no nível superior (intensional) a definição do conceito *Restaurant* nas fontes de dados DBpedia, LGD e GeoNames. Note também a relação dessas classes com classes superiores de cada fonte (*rdfs:subclassOf*) e suas relações de equivalência

(*owl:equivalentClass*). No nível inferior (extensional) da Figura 4, estão representadas duas instâncias de restaurante de fontes distintas (DBpedia e GeoNames) ligadas pela relação de iguadade (*owl:sameAs*), o que indica que se referem a mesma entidade, o restaurante *Ferrara Bakery and Cafe*.

Uma base de conhecimento é um conjunto de descrições de conceitos (conceitualizações em uma ou mais ontologias) e de instâncias. A ontologia é ao mesmo tempo dado e modelo conceitual de uma base de conhecimento. A publicação e consumo de bases de conhecimento na *Web* é orientada por um conjunto de diretrizes que permitem conexões entre recursos de diferentes fontes de dados por meio *links*, *i.e.* propriedades como *owl:sameAs* (para instâncias) e *owl:equivalentClass* (para os conceitos), como ilustrado na Figura 4. As diretrizes para a publicação de dados conectados (*Linked Data*) (BERNERS-LEE, 2006) são:

1. Usar URIs como nomes para coisas.
2. Usar URIs HTTP para que as pessoas possam procurar esses nomes.
3. Quando alguém procurar uma URI, prover informação útil, usando padrões (RDF, SPARQL).
4. Incluir *links* para outras URIs de modo que possam permitir a descoberta de mais coisas.

Os dados conectados alicerçam a *Web* de dados pela adoção de alguns padrões: um mecanismo de identificação global e único (*Uniform Resource Identifiers* - URIs), um modelo de dados comum (*Resource Description Framework* - RDF) e um protocolo e linguagem de consulta para acesso aos dados (SPA RDF *Query Language* - SPARQL).

URIs (BERNERS-LEE, 2005) são utilizadas no contexto de dados ligados para identificar unicamente conceitos, instâncias e propriedades. Ao dereferenciar uma URI, é obtida a descrição RDF do conceito, instância ou propriedade identificado. O modelo RDF (MANOLA; MILLER, 2004) é descentralizado, baseado em grafo e extensível, projetado para a representação integrada de dados de fontes diversas. Uma descrição RDF composta de declarações, como a formalmente definida pela Definição 3.

Definição 3. Uma **declaração RDF** é uma tripla $d = (s, p, o)$, onde s é o sujeito, p é o predicado e o é o objeto da declaração. Sujeito e predicado são representados por um recurso unicamente identificado

por uma URI u. O predicado p representa uma propriedade de s. O objeto é um valor de uma propriedade de s, podendo ser representado por outro recurso ou por um valor alfanumérico.

```
@prefix dbr: <http://dbpedia.org/resource/>.
@prefix gn: <http://sws.geonames.org/>.
@prefix gno: <http://www.geonames.org/ontology#>

dbr:Ferrara_Bakery_and_Cafe
    rdfs:label "Ferrara Bakery and Cafe"@en;
    geo:geometry "POINT(-73.997016906738 40.719039916992)"
^^virtrdf:Geometry;
    owl:sameAs dbr:Ferrara_Bakery_and_Cafe, gn:7288706.

gn:7288706
    gn:featurecode gno:S.REST.
```

Figura 5 – Triplas RDF, formato N3, adaptadas da DBpedia

O SPARQL é um protocolo e uma linguagem de consulta e manipulação de dados armazenados em repositórios RDF. Sua linguagem possui expressividade equivalente a da álgebra relacional e SQL (*Structured Query Language*) (ANGLES; GUTIERREZ, 2008; PERRY; SHETH; JAIN, 2008).

A Figura 5 mostra algumas triplas sobre o recurso *Ferrara*[3] obtidas a partir da consulta SPARQL ilustrada pela Figura 6. Foram selecionados os valores das propriedades *rdfs:label*, *geo:geometry*, *owl:sameAs*. Além deste recurso, a Figura 5 mostra o valor da propriedade *gn:featurecode* do recurso *gn:7288706* [4].

```
PREFIX dbr:<http://dbpedia.org/resource/>
SELECT * WHERE {
    dbr:Ferrara_Bakery_and_Cafe ?p ?o.
} LIMIT 100
```

Figura 6 – Consulta SPARQL sob DBpedia

Recursos de KB, quando referenciados como informação sobreposta em um conjunto de anotações semânticas, caracterizam-se como

[3] *dbr:Ferrara_Bakery_and_Cafe*, dbr é prefixo para http://dbpedia.org/resource/
[4] gn é prefixo para http://sws.geonames.org/

recursos de enriquecimento semântico (Definição 4). Esta definição limita-se apenas a MDSs semanticamente anotados neste trabalho, embora estenda-se a qualquer tipo de dado alvo de uma anotação semântica.

Definição 4. Um **recurso de enriquecimento semântico** (*Semantically Enriching Resource*) é um URI res que referencia um conceito ou uma instância de base de conhecimento, o qual foi utilizado como valor de uma anotação semântica $a = (mds, rel, res)$.

Por exemplo, o conceito de restaurante e/ou instância específica de restaurante podem ser utilizados como valor de uma anotação de um MDS para indicar um local visitado em tal segmento de movimento. No exemplo da subseção anterior, *lgd:way73321157* é o recurso de enriquecimento semântico utilizado na associação $a''' = (mds'', visits, lgd{:}way73321157)$.

2.4 DATA WAREHOUSING

As técnicas de *Data Warehousing* (KIMBALL, 1996) permitem gerir e reorganizar vasta quantidade de dados relativos a um determinado fenômeno (*e.g.* condições metereológicas, de negócio) em *Data Warehouses* (DWs), *i.e.* bases de dados multidimensionais, com o intuito de realizar análises e predições a respeito deste fenômeno. O DW é uma coleção de dados que apresenta as seguintes características:

- Integrada – composta de dados provenientes de diferentes fontes (*e.g.*, sistemas transacionais, fontes externas);
- Orientada a um assunto – formada com o intuito de resolver um problema específico (análise de fenômeno);
- Variável no tempo – que contém dados que compreendem um laço temporal mais extenso que coleções de dados normalmente memorizados em sistemas operacionais;
- Não volátil – cuja informação armazenada é estática.

O modelo multidimensional de DWs organiza dados em fatos e dimensões de análise, com o intuito de produzir um *hipercubo de dados* (*data cube*), cuja cada célula possui medidas de interesse.

Por exemplo, a Figura 7 ilustra um hipercubo de dados sobre *tweets*, organizados pelas dimensões de análise *spatial object* (*e.g.*, *Bairro::Agronômica*), *spatial concept* (*e.g.*, *Supermarket*) e *timestamp*

(*e.g.*, *11/10/2014*). Um fato corresponde, por exemplo, a medidas de *tweets* postados num dia particular do ano, num lugar e num tipo de lugar específicos. Um exemplo das medidas que podem ser colocadas em cada célula deste cubo é o número de *tweets* postados (*e.g.*, *qty*). Além disso, a Figura 7 também mostra que há apenas 2 *tweets* postados em supermercados do bairro Agronômica no dia 11 de Outubro de 2014.

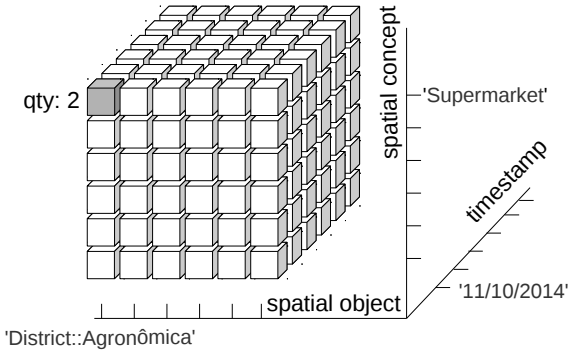


Figura 7 – Representação gráfica de um hipercubo de dados

Dimensões de análise são hierarquias para organização e análise dos fatos em diversos níveis de abstração (CABIBBO; TORLONE, 1998). Por exemplo, a Figura 8 ilustra uma hierarquia de dimensão com os níveis: *Spatial Object* (lugares), *District* (bairro), *City* (cidade), *Country* (país). Esta hierarquia agrupa lugares em níveis administrativos de acordo com a contenção espacial. As setas entre membros de níveis da dimensão (*e.g.*, *Angeloni*→*Agronômica*→*Florianópolis*) representam a relação de estar contido (*part of*). A raiz desta hierarquia é *Earth*, pois todos os *Spatial Object* estão contídos na Terra.

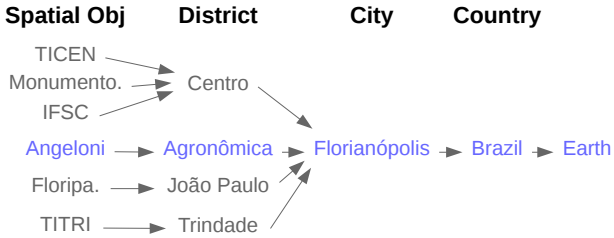


Figura 8 – Exemplo de hierarquia sobre a dimensão *Spatial Object*

A representação gráfica de DWs torna-se mais complexa a medida que aumenta o número de dimensões de análise. A modelagem conceitual, como por meio do modelo de fato dimensional (*Dimensional Fact Model* - DFM) (GOLFARELLI; MAIO; RIZZI, 1998), é um fundamento necessário para a construção de um DW bem documentado e que satisfaça os requisitos específicos da análise de determinado fenômeno.

A Figura 9 ilustra o esquema de fato (*fact scheme*) *Movement Segment*, representado por DFM. Este esquema descreve o movimento de objetos móveis, onde o fato é representado por uma caixa rotulada pelo nome do fato (*Movement Segment*) e, tipicamente, com uma ou mais medidas de fato (*episodyQty*, *elapsedTime*, *distanceTravelled*). Dimensões são representadas por círculos diretamente conectados ao fato (*i.e. M.O.*, *spatial concept*, *spatial object*, *timestamp*), e os círculos remanescentes são atributos das dimensões. Atributos não-dimensionais são sempre terminais e são representados por linhas (*e.g.*, *address*, *category*).

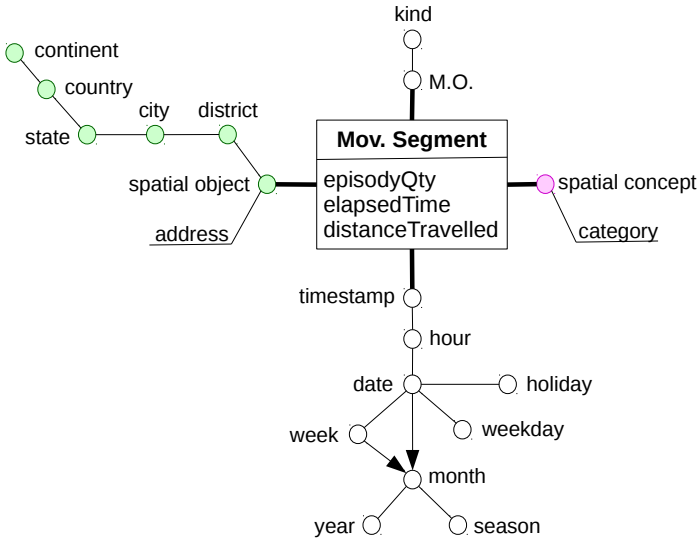


Figura 9 – O esquema de fato *Movement Segment*

No DFM, cada dimensão é composta por uma hierarquia de atributos, cuja aresta direcional (de sentido: fato, dimensão, atributos) entre atributos representa uma relação de cardinalidade *-para-um* (*e.g.*, há uma relação *muitos-para-um* entre *city* e *state*). A relação

de cardinalidade *-para-um* entre atributos de dimensão é respeitada pelos valores que estes atributos assumem (*e.g.*, o *country* (país) de Agronômica (valor de *district*) é Florianópolis, *i.e.* Agronômica.*city* = Florianópolis).

É possível representar o hipercubo de dados também utilizando o modelo relacional. A Figura 10 apresenta um exemplo de esquema relacional de um DW. A tabela *FactMO* armazena fatos no formato de tuplas – idMO, idSpatialC, idSpatialO, idTime, episodyQty, elapsedTime, distanceTravelled – onde os 4 primeiros atributos forma o identificador primário e os três últimos são medidas de fatos. A tabela *SpatialC*, *SpatialO* e *Time* armazenam dimensões de análise descritas por tuplas contendo o identificador (*e.g.*, *idTime*) e atributos de dimensão (*e.g.*, *year*, *month*, *day*, *hour*). Neste DW, a dimensão *SpatialC* apresenta uma hierarquia de atributos de um único nível. A tabela *SpatialGeom* armazena geometrias de lugares referenciados pela dimensão *SpatialO*.

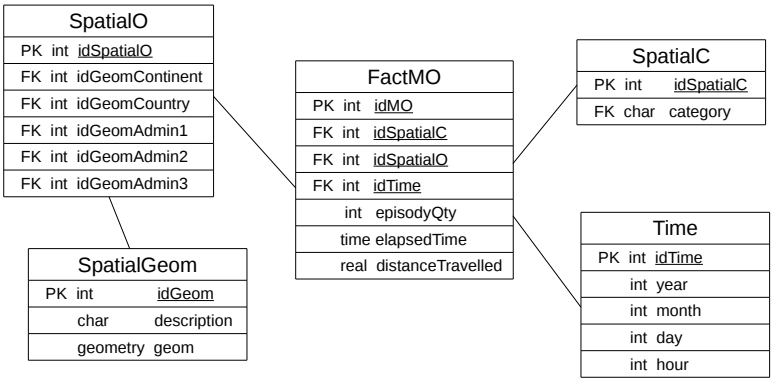


Figura 10 – Esquema relacional do DW

Os dados de DWs podem provir de diversas fontes, as quais incluem sistemas legados e a própria *Web*. Tais dados precisam passar por um proceso de Extração, Transformação e Carga (*Extraction, Transformation and Loading* - ETL) para serem acomodados em um esquema dimensional de um DW. Uma vez acomodados no DW, os dados podem ser analisados através do OLAP (On-Line Analytical Processing). O processo OLAP é suportado por um conjunto de ferramentas que oferecem implementações eficientes de algoritmos para realizar operadores como *drill-down*, *roll-up*, *slice* e *dice*, os quais permitem avaliar medidas contidas nas tabelas fato em diversos níveis de abstração e foco, de

acordo com as dimensões de análise.

As iniciativas da *Web 2.0* e *3.0* (também chamada *Web* de Dados e *Web* Semântica) tornaram acessíveis grandes quantidades de dados semiestruturados (*e.g.*, *Extensible Markup Language* - XML, RDF) e não estruturados (*e.g.*, textos livres de corpos de email, posts em mídias sociais), importantes ou convenientes para a análise de dados em DWs (INMON; STRAUSS; NEUSHLOSS, 2008).

Por exemplo, considere que o DW ilustrado pela figura 7 deve analisar os lugares onde houveram *tweets*, durante os eventos esportivos de 2014, nos diferentes bairros e cidades da região de *Grande Florianópolis*, no estado de Santa Catarina, Brasil. A análise deve considerar a data e hora de postagem dos *tweets*, localização, e categoria de lugar, onde cada uma dessas dimensões de análise devem ser organizadas em hierarquias de níveis bem definidos, *e.g.*, *Dia*→*Semana*→*Mês*→*Ano*, *Endereço*→*Bairro*→*Cidade*→*Região*, e *Subsubcategoria*→*Subcategoria* →*Categoria*. Para isto, o DW deve integrar dados de *tweets* (coletado em uma base NO-SQL a partir da API da mídia social Twitter) com dados geográficos sobre lugares (coletados de LODs em formato RDF), lista de eventos esportivos de 2014 (em formato não estruturado coletado de blogs e wikis esportivas), e finalmente de ontologias descrevendo a hierarquia de contenção geográfica e a hierarquia de categorias de lugar.

As tecnologias da *Web* Semântica têm sido aplicadas de diferentes modos para suportar análise em DW (PARDILLO; MAZÓN, 2011). Em Abelló et al. (2015), são definidos 5 critérios de categorização de sistemas OLAP:

- Materialização – que considera o nível de materialização dos dados integrados: completo, parcial, armazenamento de resultado e virtual.

- Transformações – que considera a complexidade das transformações: complexa, tolerante a partição, leve.

- Atualização – que considera com qual frequência é realizada a integração de dados: periódica, microlotes, sob demanda, *Right-time*, fluxo de dados.

- Estruturação – que considera a estrutura das fontes de dados: estruturada, semiestruturada, não estruturada.

- Extensibilidade – que considera quão dinâmico é o conjunto de fonte de dados de entrada: estático, evolução, dinâmico.

DWs tradicionais caracterizam-se pela materialização completa, transformações complexas, atualização periódica, fontes de dados estruturadas e extensibilidade estática.

Situados entre os DWs tradicional e exploratório, os DWs semântico-conscientes (*Semantic-aware* ou *Semantic-enable*) aplicam tecnologias da *Web* Semântica para satisfazer requisitos de DW tradicionais. Estes sistemas necessitam explorar fontes de dados semiestruturados ou não estruturados (*Web 2.0* e *3.0*) não necessariamente estáticos, sem utilizar materialização virtual e atualização sob demanda.

Dos usos de tecnologias da *Web* Semântica em DWs, a análise de dados semanticamente anotados por recursos que referenciam ontologias seguem diferentes abordagens. A primeira mapeia hierarquias de recursos de enriquecimento semântico em hierarquias de dimensão com níveis bem definidos para formar dimensões semânticas (*semantic dimensions*) (ANDERLIK; NEUMAYR; SCHREFL, 2012; NEBOT; BERLANGA, 2012). A segunda não transforma a ontologia em modelo relacional e realiza a análise dos dados triplificados (KäMPGEN; HARTH, 2011; ETCHEVERRY; VAISMAN; ZIMÁNYI, 2014).

2.5 CONSIDERAÇÕES FINAIS

Este capítulo apresentou fundamentos básicos sobre dados sobre movimento, anotações, *Web* semântica, LOD e *data warehouses*. Embora já existam trabalhos na literatura que mapeiam hierarquias de recursos de enriquecimento semântico em hierarquias de dimensão com níveis bem definidos para formar dimensões semânticas (ANDERLIK; NEUMAYR; SCHREFL, 2012; NEBOT; BERLANGA, 2012), pouco é discutido sobre a extração e adaptação de hierarquias de recursos fontes de dados como KBs e LOD.

Esta dissertação contribui com um método que propõe: criar anotações semânticas; extrair hierarquias de recursos de diversas fontes de dados a partir das anotações semânticas criadas; e adaptar as hierarquias de recursos extraídas por meio de algoritmos automatizados e da edição manual. Este método, assim como suas definições básicas e exemplos de utilização, no capítulo a seguir.

3 ADAPTAÇÃO DE HIERARQUIAS DE RECURSOS

Este capítulo apresenta a contribuição desta dissertação, um método para a geração automatizada de dimensões de análise a partir de conjuntos de dados semanticamente anotados e por meio da adaptação de hierarquias de recursos, ilustrado pela Figura 11.

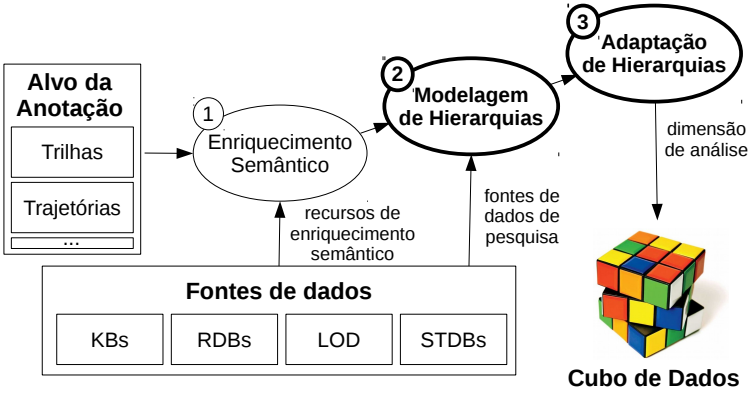


Figura 11 – Modelo geral do método de adaptação de hierarquias de recursos

O método proposto é dividido em 3 fases: *Enriquecimento Semântico*, *Modelagem de Hierarquias* e *Adaptação de Hierarquias*. A fase de *Enriquecimento Semântico* tem o objetivo de produzir conjuntos de dados semanticamente anotados. No estudo de caso desta dissertação, os alvos da anotação são dados sobre movimento (trajetórias ou trilhas) e o recursos de enriquecimento semântico podem ser extraídos de bases de conhecimento (KBs), dados abertos conectados (LOD), bases de dados relacionais (RDBs) ou bases de dados espaço-temporais (STDB). Esta fase não é o enfoque desta dissertação e é melhor detalhada em diversos trabalhos da literatura (YAN et al., 2013; RINZIVILLO et al., 2013; FILETO et al., 2013; MAY; FILETO, 2014).

A fase de *Modelagem de Hierarquias* objetiva a extração de hierarquias de recursos de diversas fontes de dados (utilizando um processo automatizado e configurável em LOD e KBs). A hierarquia de recursos extraída é composta de recursos conectados a recursos de enriquecimento semântico por meio de propriedades de equivalência (*e.g.*, *owl:sameAs*, *owl:equivalentClass*) ou que expressem relações de

ordenação parcial (*e.g.*, *is a* e *part of*). A fase de *Adaptação de Hierarquias* objetiva a adequação de hierarquias de recursos para viabilizar seu uso como dimensão de análise (utilizando um processo automatizado e configurável para reduzir o número de recursos da hierarquia).

Primeiro, este capítulo apresenta definições básicas sobre conjuntos de dados sobre movimento semanticamente anotados (SMoD), hierarquia de recursos e contagem de associações. Depois, este capítulo descreve cada uma das fases do método proposto. Por fim, este capítulo apresenta considerações finais.

3.1 DEFINIÇÕES BÁSICAS

Um conjunto de dados sobre movimento semanticamente anotados (*Semantically annotated Movement Dataset* - SMoD) (Definição 5) é um conjunto de segmentos de dados sobre movimento (descritos pela Definição 1), associados a recursos de enriquecimento semântico (Definição 4) por anotações semânticas não-intrusivas. Uma anotação semântica pode ser representada por uma associação da forma (mds, rel, at), como descrita na Definição 2, onde rel é uma relação semântica e o atributo temático at é um recurso de enriquecimento semântico identificando um objeto ou conceito em uma base de conhecimento.

Definição 5. Um **conjunto de dados sobre movimento semanticamente anotados** é uma tupla $SMoD = (MoD, R, A)$, onde MoD é um conjunto de MDSs, R é um conjunto de recursos de enriquecimento semântico e $A = \{at \in MoD \times R\}$ é um conjunto de anotações semânticas associando MDSs a recursos.

Por exemplo, a Figura 12 ilustra um SMoD cujo conjunto de MDSs é uma coleção de *tweets*, os recursos de enriquecimento semântico são representados por elipses verdes, e as anotações semânticas são representadas por arestas. Recursos de enriquecimento semântico deste SMoD são conectados a outros recursos por meio de propriedades que definem: relações de ordenamento parcial topológico (*i.e. partOf* ou *contains*) *gn:parentADM2* e *gn:parentFeature*, e relações de ordenamento parcial de subsunção (*i.e. isA*) *rdf:type* e *rdfs:subClassOf*).

Relações de ordenamento parcial proporcionam a representação de hierarquia de recursos, como a hierarquia formalmente definida pela Definição 6.

Definição 6. Uma **hierarquia de recursos** (*Resource Hierarchy*) é um digrafo acíclico e fracamente conexo (*weakly connected directed*

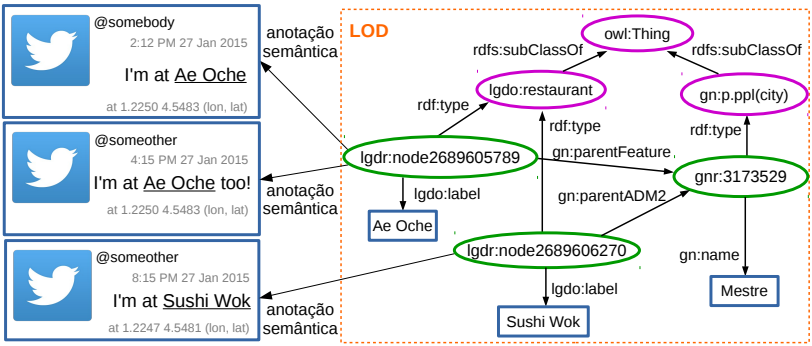


Figura 12 – Associações entre *tweets* e recursos de LOD

acyclic graph - DAG) $H = (R, P)$, onde o conjunto de vértices R é um conjunto de recursos identificados por URIs, o conjunto de arestas $P = (u_i, u_j)$ com $i, j \in \mathbb{Z}$ e $u_i, u_j \in R$ e P é um conjunto de relações semânticas entre recursos em R definidas por uma propriedade de ordenamento parcial (*e.g.*, *is a* ou *subClassOf*, *part of* ou *contained*, *type*).

Dado um conjunto de dados semanticamente anotados e uma hierarquia de recursos, podemos classificar a relação entre um dado alvo de anotação e um recurso da hierarquia como direta e indireta. Um dado alvo de anotação t e um recurso da hierarquia r possuem uma relação direta quando existe uma anotação semântica em forma de associação (Definição 2) (t, rel, r), *i.e.* quando r for um recurso de enriquecimento semântico de t. Por outro lado, um dado alvo de anotação t e um recurso da hierarquia r possuem uma relação indireta se e somente se existe uma cadeia de recursos $\{d, ..., r\}$ conectados por propriedades RDF, tal que existe relação direta entre t e d.

Por exemplo, na Figura 12 o recurso *lgdr:node2689605789* (restaurante *Ae Oche*) apresenta 2 relações diretas (primeiro e segundo *tweet*), e o recurso *lgdr:node2689606270* (restaurante *Sushi Wok*) apresenta 1 relação direta (terceiro *tweet*), onde cada relação direta é representada por uma aresta *anotação semântica*. O recurso *gn:p.ppl(city)* está indiretamente relacionado a todos os 3 *tweets* através das cadeias de recursos: {*lgdr:node2689605789*, *gnr:3173529*, *gn:p.ppl(city)*} e {*lgdr:node2689606270*, *gnr:3173529*, *gn:p.ppl(city)*}.

Finalmente, dado um conjunto de dados semanticamente anotados e uma hierarquia de recursos, podemos definir a frequência de uso de um recurso em anotações semânticas, considerando relações diretas e indiretas, pela Definição 7.

Definição 7. Seja T um conjunto de dados alvo de anotação, R um conjunto de recursos, $N \subseteq M \times R$ um conjunto de anotações semânticas entre M e R, $C \subseteq R \times R$ um conjunto de conexões entre recursos, e $A = N \cup C$ um conjunto de associações. A **frequência de uso** (*use frequency, number of hits*) $h(r, A)$ de um recurso $r \in R$ com respeito a A é o número de dados alvo de anotação distintos em $t \in T$ tal que exista relação direta ou indireta entre r e t.

Por exemplo, o recurso *gn:p.ppl(city)* possui 0 hits diretos, pois não existe relação direta entre ele e qualquer *tweet*. Porém, a frequência de uso de *gn:p.ppl(city)* é 3, pois existe uma relação indireta entre este recurso e cada um dos *tweets*.

3.2 MÉTODO PARA A ADAPTAÇÃO DE HIERARQUIAS

Esta dissertação propõe o método para adaptação de hierarquias, ilustrado pela Figura 13, com o objetivo de gerar de modo automatizado dimensões de análise para conjuntos de dados semanticamente anotados. Este método deve ser acrescentado ao processo de ETL convencional para permitir a construção de DWs que analisem dados semanticamente anotados.

O método proposto é dividido em 3 fases: *Enriquecimento Semântico*, *Modelagem de Hierarquias* e *Adaptação de Hierarquias*. A fase de *Enriquecimento semântico* não é o enfoque desta dissertação e é melhor detalhada em diversos trabalhos da literatura (YAN et al., 2013; RINZIVILLO et al., 2013; FILETO et al., 2013; MAY; FILETO, 2014). Portanto, este trabalho tem enfoque nas fases de *Modelagem de Hierarquias* e *Adaptação de Hierarquias* (em negrito). O método proposto possui 9 passos, igualmente distribuídos pelas 3 fases. As fases e seus passos são descritos as seguir.

3.2.1 Enriquecimento Semântico

A fase de *Enriquecimento Semântico* tem o objetivo de produzir conjuntos de dados semanticamente anotados. No estudo de caso desta dissertação, os alvos da anotação são dados sobre movimento (trajetórias ou trilhas) e o recursos de enriquecimento semântico são extraídos de coleções de dados abertos conectados (LOD).

De modo geral, recursos de enriquecimento semântico também podem ser extraídos de bases de conhecimentos (KBs) de forma análoga

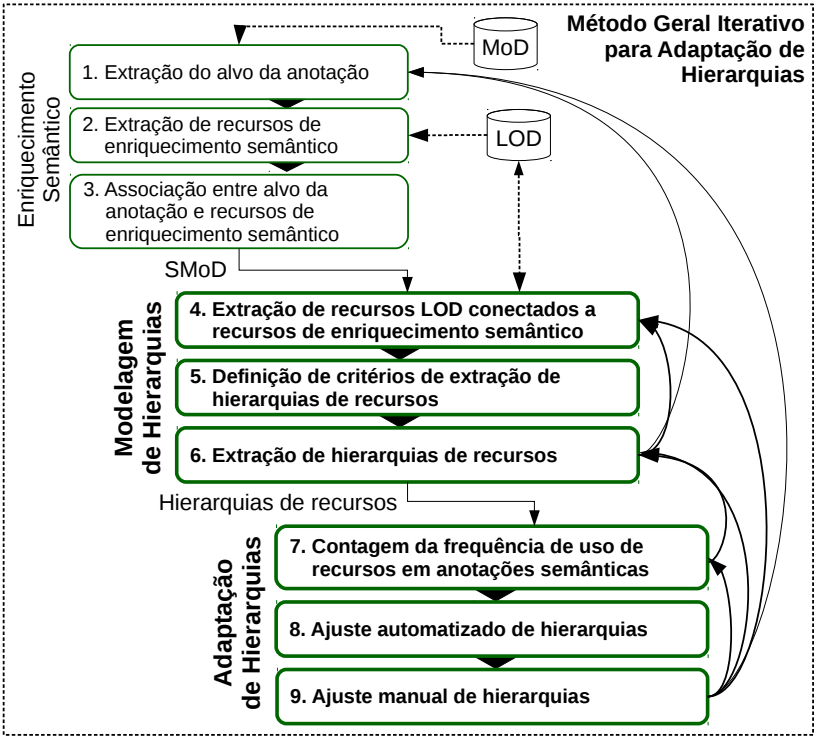


Figura 13 – Método iterativo para a adaptação de hierarquias de recursos

aos de coleções de LOD, ou extraídos de bases de dados relacionais (RDBs) e espaço-temporais (STDBs) se aplicadas adaptações, *i.e.* desde que estes dados sejam representados por recursos através de um processo de triplificação.

A fase de *Enriquecimento Semântico* possui os seguintes passos:

Passo 1 - Extração / obtenção dos dados alvos de anotações semânticas. Os dados alvos de anotação e informações necessárias para criar associações devem permitir acesso de leitura.

Passo 2 - Extração / obtenção de recursos de enriquecimento semântico. Os recursos de enriquecimento semântico e informações necessárias para criar associações devem permitir acesso de leitura.

Passo 3 - Associações entre dados alvos da anotação e recursos de enriquecimento semântico são criadas. Diversos algoritmos são

propostos na literatura para criar anotações semânticas. Além disso, anotações semânticas podem ser criadas para expressar diferentes informações (*e.g.*, local visitado, meios de transporte, condições do ambiente, objetivos e atividades do MO (BOGORNY et al., 2014).

Por exemplo, no estudo de caso desta dissertação, o Passo 1 foi executado por um extrator que coletou informações de *tweets* como usuário, conteúdo textual e posição geográfica de postagem e as armazenou em uma RBD. O Passo 2 foi executado por um extrator de recursos de coleções de LOD, que extraiu informações de recursos do DBpedia e LinkedGeoData (LGD) como URI, nome do local e posição geográfica e as armazenou em uma RDB. Por fim, Passo 3 executou um algoritmo para criar associações considerando proximidade espacial e similaridade textual de *tweets* e recursos (MAY; FILETO, 2014), como ilustrado na Figura 14.

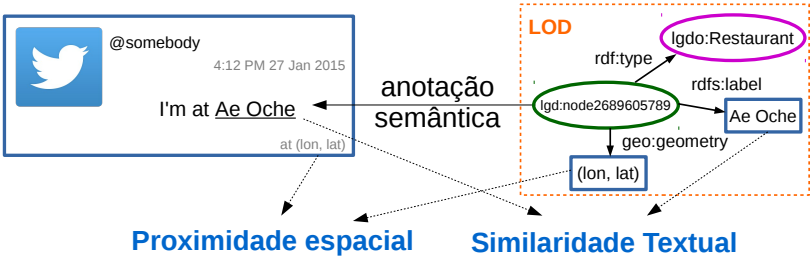


Figura 14 – Anotação semântica de *tweet* considerando proximidade espacial e similaridade textual

As associações entre *tweets* e recursos do DBpedia e LGD são anotações semânticas que expressam o local visitado pelo usuário durante a postagem do *tweet*. O conjunto de dados semanticamente anotados gerado é representado como SMoD descrito pela Definição 5, e serve como dado de entrada para a fase de *Modelagem de Hierarquias*.

3.2.2 Modelagem de Hierarquias

A fase de *Modelagem de Hierarquias* objetiva a extração de hierarquias de recursos conectados aos recursos de enriquecimento semântico. Hierarquias de recursos podem ser extraídas por meio de um processo automatizado e configurável. Este processo tem como entrada:

as fontes de dados de pesquisa, as conexões a serem exploradas, e os critérios de extração de hierarquias.

A extração de recursos é restringida por fontes de dados de pesquisa (coleções de DAC como *DBpedia*, *LinkedGeoData*, e/ou BCs do usuário do processo) selecionadas pelo usuário do processo, *i.e.* recursos externos às fontes de dados de pesquisa são ignorados pela extração. As conexões entre recursos exploradas pela extração expressam relacionamentos de equivalência (*e.g.*, *owl:sameAs*, *owl:equivalentClass*) ou de ordenação parcial como *is a* (*e.g.*, *rdf:type* e *rdfs:subClassOf*) e *part of* (*e.g.*, *gn:parentFeature* e *gn:parentCountry*). Fontes de dados de recursos extraídos podem ser diferentes das fontes de dados dos recursos de enriquecimento semântico, desde que conexões de recursos entre as diferentes fontes estejam disponíveis nas fontes de dados de pesquisa.

Os recursos extraídos são usados para compor a hierarquia de recursos, respeitando os critérios de extração de hierarquias. Por exemplo, em nossos experimentos foram definidos de dois critérios: um de mapeamento direto de conexões exploradas, outro de exploração de cadeias de recursos. O critério de mapeamento direto define uma lista de conexões exploradas, ordenadas pela prioridade de preenchimento, para preencher determinado nível da hierarquia.

Algumas conexões são utilizadas para indicar recursos de diferentes níveis da hierarquia (*e.g.*, *rdfs:subClassOf*, *gn:parentFeature*), impossibilitando o mapeamento direto. Para estas conexões, são extraídas cadeias de recursos. Nestas cadeias, são eliminadas cadeias mais curtas até determinado recurso (conexões antecipadas) e é selecionada uma única cadeia para cada recurso de enriquecimento semântico de acordo com o critério de extração de hierarquias por exploração de cadeias de recursos definida pelo usuário (*e.g.*, escolha da cadeia mais longa). Por fim é adicionado um ancestral comum para todas as cadeias extraídas (*e.g.*, o recurso *owl:Thing*).

A cada recurso de enriquecimento semântico (obtidos da fase de *Enriquecimento Semântico*) e a cada recurso encontrado na extração, é aplicado o Passo 4:

Passo 4 - O recurso explorado é dereferenciado em busca de conexões da lista de conexões exploradas. Recursos não pertencentes às fontes de dados de pesquisa selecionadas pelo usuário do processo são ignorados. A lista de conexões exploradas e fontes de dados de pesquisa são previamente informadas pelo usuário do processo.

Após a exploração de recursos e conexões (Passo 4), os dados são armazenados em um repositório de triplas. Durante o Passo 5, o usuário

do método define os critérios de extração de hierarquias. O Passo 6 utiliza esses critérios para selecionar uma única cadeia de recursos, para cada recursos de enriquecimento semântico, que irá compor a hierarquia de recursos.

Passo 5 - Definição de critérios de extração de hierarquias. O usuário pode definir critérios de mapeamento direto, de exploração de cadeias e a prioridade de escolha entre a cadeia extraídas por mapeamento direto e as cadeias extraídas por exploração de cadeias.

Passo 6 - Extração de hierarquias de recursos. Para cada recurso de enriquecimento semântico, é construída uma cadeia de recursos por meio de conexões de mapeamento direto e são extraídas cadeias de recursos por meio de conexões de exploração de cadeias. Depois, é selecionada apenas uma cadeia de recursos para cada recurso de enriquecimento semântico, considerando os critérios de extração de hierarquias definidos durante o Passo 5.

Por exemplo, a Figura 15 apresenta um SMoD de *tweets* anotados com recursos sobre lugares de interesse (PoIs) expressando o local visitado. As associações entre *tweets* e recursos foram criadas considerando a proximidade espacial e similaridade textual (MAY; FILETO, 2014), durante a fase de *Enriquecimento Semântico*.

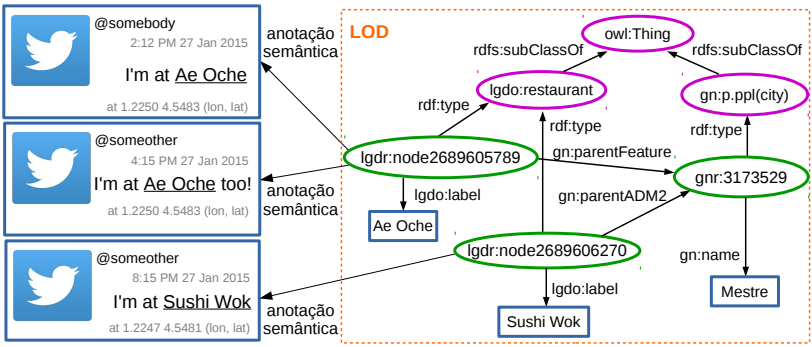


Figura 15 – Associações entre *tweets* e recursos de LOD

Durante a fase de *Modelagem de Hierarquias*, no Passo 4, recursos de enriquecimento semântico (*e.g.*, *lgdr:node2689605789*, *lgdr:node2689606270*) são explorados nas fontes de dados de pesquisa em busca de recursos sobre a mesma entidade. Depois, estes recursos são verificados em busca de conexões declaradas na lista de conexões exploradas (*e.g.*, *rdf:type*, *rdfs:subClassOf*, *gn:parentADM2*, *gn:parentFeature*).

Os recursos conectados por conexões exploradas (*e.g.*, *gnr:3173529*, *lgdo:restaurant*) são posteriormente explorados pelo Passos 4. Recursos e conexões explorados são armazenados em um repositório de triplas.

Após a exploração de todos os recursos e conexões, o usuário define os critérios de extração de hierarquias com base nos dados extraídos durante o Passo 5. Para hierarquias de recursos de níveis expressos por conceitos bem definidos, é sugerido o critério de extração de hierarquias por mapeamento direto (*i.e.* para cada nível é definida uma lista de conexões ordenada segundo a preferência de preenchimento do nível em questão com o valor da conexão explorada). Por outro lado, para hierarquia de recursos de níveis sem conceitos bem definidos, é sugerido o critério de extração de hierarquias por exploração de cadeia. Este critério define a preferência por qual das cadeias de recursos extraídas (*e.g.*, optar sempre pela cadeia de recursos mais longa).

Para cada recurso de enriquecimento semântico, o Passo 6 utiliza os critérios de extração de hierarquias por mapeamento direto para construir cadeias de recursos preenchendo cada nível da hierarquia. Depois, cadeias de recursos são extraídas considerando as conexões de exploração de cadeias. O Passo 6 elimina cadeias de recursos com conexões antecipadas. O Passo 6 seleciona uma única cadeia para aquele recurso de enriquecimento semântico de acordo com os critérios de extração de hierarquias definidos pelo usuário.

As cadeias de recursos selecionadas compõem uma hierarquia de recursos assim como a descrita pela Definição 6, e serve como dado de entrada para a fase de *Adaptação de Hierarquias*.

3.2.3 Adaptação de Hierarquias

A fase de *Adaptação de Hierarquias* objetiva a adequação de hierarquias de recursos para viabilizar seu uso na análise de informação. Diversos algoritmos podem ser aplicados as hierarquias obtidas na fase de *Modelagem de Hierarquias* com diversos intuitos (*e.g.*, redução de número de recursos pela eliminação ou agregação de recursos, adição de conceitos ou instâncias à hierarquia). Esta fase consiste nos seguintes passos:

Passo 7 Análise estatística de hierarquias de recursos. Dados estatísticos como a frequência de uso, descrita pela definição 7, são gerado e armazenados.

Passo 8 Adaptação automatizada de hierarquias de recursos. Algorit-

mos de adaptação aplicam correções às hierarquias utilizando os dados estatísticos anteriormente gerados.

Passo 9 Adaptação manual de hierarquias de recursos. O usuário do método aplica correções manuais a hierarquia de recursos extraída e adaptada por processos automatizados e configuráveis.

No estudo de caso desta dissertação, durante o passo 7, foram geradas as frequências de uso de cada recurso das hierarquias. No Passo 8, o algoritmo para adaptação automatizada de hierarquias aplicado foi o *SimpleTailoring*. O *SimpleTailoring* tem o objetivo de esconder recursos cuja frequência de uso seja menor que o número estipulado pelo usuário do algoritmo. O Passo 9 é executado com o auxílio de ferramentas de dedição de ontologias, visto que no estudo de caso, hierarquias de recursos foram descritas por triplas RDF.

O algoritmo *SimpleTailoring*, utilizado pelo estudo de caso desta dissertação, é descrito a seguir.

3.2.4 Algoritmo para adaptação automatizada de hierarquias

SimpleTailoring (Algoritmo 1), aplicado no estudo de caso desta dissertação durante o Passo 8 da fase de *Adaptação de Hierarquias*, tem o objetivo de esconder recursos cuja frequência de uso (Definição 7) seja menor que o número estipulado pelo usuário do algoritmo. Recursos de mesmo pai escondidos deste modo são agredados em um único nodo da hierarquia.

O algoritmo *SimpleTailoring* tem como entrada uma hierarquia de recursos $\mathcal{H}$, as frequências de uso de cada recurso da hierarquia obtidas pelo Passo 7 e um limiar inferior σ estipulado pelo usuário. O algoritmo procede pelos seguintes passos:

1. Para cada recurso de $\mathcal{H}$, a função *Filter* (Algoritmo 2) verifica se a frequência de uso do recurso é maior ou igual a σ. Caso não satisfazer o limiar, o rótulo deste recurso é substituído por *"Other"*. Recursos rotulados desta maneira são considerados não-relevantes para a hierarquia adaptada.

2. Depois da função *Filter*, um recurso de $\mathcal{H}$ pode ter vários nodos filhos rotulados como "Other". A função *Merge* (Algoritmo 4) agrega os nodos "Other"de um mesmo nodo pai em um único nodo "Others".

3. Enfim, a função View (Algoritmo 3) encontra o menor ancestral comum (*i.e.* o nodo de menor nível que seja pai de todos os nodos do nível inferior). O menor ancestral comum é definido como a raiz da hierarquia adaptada, eliminando nodos ancestrais a ele da nova hierarquia.

Algorithm 1: SimpleTailoring($\mathcal{H}$,*hits*,σ)

Input: Resource hierarchy $\mathcal{H}$, use frequency of each resource *hits*, threshold σ
Output: adapted hierarchy $\mathcal{A}$

```
1 A ← H;
2 Filter(A, hits, σ);
3 Merge(A.root, hits);
4 A.root ← View(A.root);
5 return A;
```

Algorithm 2: Filter($\mathcal{A}$,*hits*,σ)

Input: Resource hierarchy $\mathcal{A}$, use frequency of each resource *hits*, threshold σ
Output: none

```
1 for each r ∈ A do
2     if r.hits < σ then
3         r.label ← "Other";
```

Algorithm 3: View(r)

Input: hierarchy root r
Output: new hierarchy root

```
1 if |r.children()| = 1 then
2     child ← unique child of r;
3     return View(child);
4 else
5     return r
```

Algorithm 4: Merge(r,*hits*)

Input: hierarchy root r, use frequency of each resource *hits*
Output: none

```
1  o ← new node;
2  o.label ← "Others";
3  o.hits ← 0;
4  for each
   child ∈ r.children() do
5      if child.label = "Other"
       then
6          o.addChildren(
7              child.children());
8          o.hits ←
9              o.hits +
               child.hits;
10         r.removeChild(child);
11     else
12         Merge(child, hits);
13 if |o.children()| > 0 then
14     r.addChild(o);
15     Merge(o, hits);
```

Obs.: A função Filter do Apêndice A, além da funcionalidade explicada nesta seção, também contabiliza a frequência de uso. Aqui este cálculo é omitido por ser feito anteriormente, no Passo 7.

Por exemplo, considere a execução do algoritmo com os seguintes parâmetros: a hierarquia de recursos e a frequência de uso de cada recurso, ilustrados pela Figura 16, e o limiar inferior σ de valor 20.

Primeiro, a função *Filter* identifica que os recursos *Santana*, *Mocca*, *Vila Mariana*, *Carrefour* (que contém respectivamente as frequências de uso 15, 1, 2, 11) não satisfazem o limiar σ (tem frequência de uso inferior a 20). Os rótulos destes recursos são substituídos por *"Other"*, como ilustrado pela Figura 17.

Depois, a função *Merge* agrega os nodos "Other"de um mesmo nodo pai em um único nodo "Others", como ilustrado pela Figura 18.

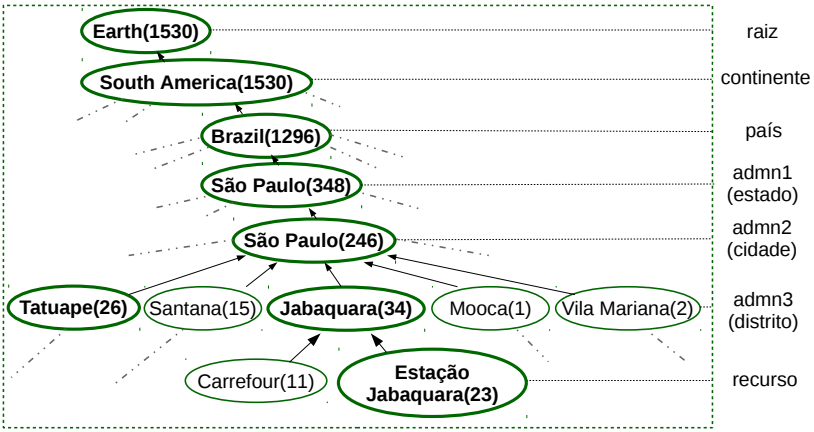


Figura 16 – Exemplo 1 da aplicação do algoritmo SimpleTailoring - entrada de dado

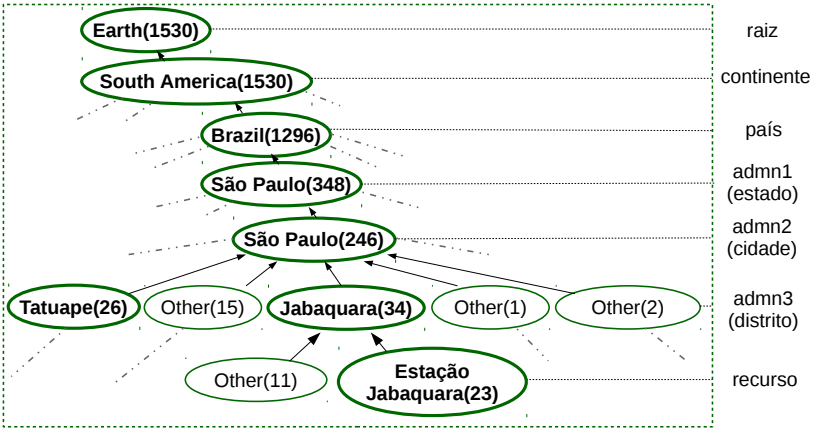


Figura 17 – Exemplo 2 da aplicação do algoritmo SimpleTailoring - omiss ão de recursos

Os recursos *Jabaquara* e *São Paulo* (nível *admin2*) possuem nodos filhos rotulado como *"Other"*. Para o recurso *Jabaquara*, a função *Merge* cria um nodo filho rotulado como "Others", adiciona os filhos do nodo "Other" referente ao recurso *Carrefour* e adiciona a frequência de uso do novo nodo "Others" a frequência de uso do nodo "Other" referente ao recurso *Carrefour* $(0 + 11 = 11)$. Para o recurso *São Paulo* (nível *admin2*), também é criado um nodo "Others", ao qual é adi-

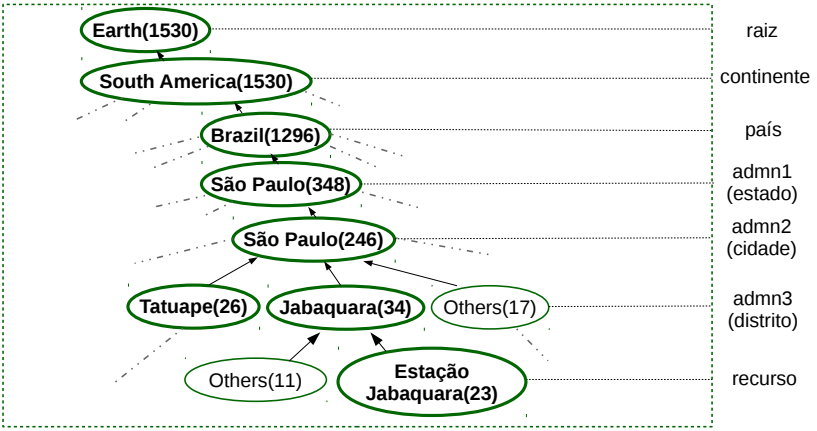


Figura 18 – Exemplo 3 da aplicação do algoritmo SimpleTailoring - agregação de recursos

cionado os filhos dos nodos "Other" referentes aos recursos *Santana*, *Mocca* e *Vila Mariana*, e a frequência de uso destes é incrementada ao daquele do mesmo modo ($0 + 15 + 1 + 2 = 17$).

Por fim, a função *View* encontra o menor ancestral comum (o recurso *South America*) e elimina seus ancestrais (o recurso *Earth*) da hierarquia adaptada.

3.3 CONSIDERAÇÕES FINAIS

O método proposto permite que resultados parciais (*i.e.* dados alvo da anotação, recursos de enriquecimento semântico, anotações semânticas, recursos e conexões extraídos, hierarquia de recursos) sejam armazenados e reutilizados em futuras iterações para aprimorar a hierarquia de recursos adaptada. Entretanto, hierarquias de recursos geradas pelo reuso de resultados parciais muito antigos podem estar desatualizadas em relação as fontes de dados de pesquisa.

As hierarquias de recursos extraídas são dependentes do conjunto de anotações semânticas de entrada e dos recursos e conexões disponíveis nas fontes de dados de pesquisa (KBs, LOD). O desenvolvimento de técnicas de enriquecimento semântico e de manutenção de KBs e LOD são desafios de pesquisa fundamentais para a viabilidade do método proposto.

A adaptação de hierarquias por redução tem o objetivo de ob-

ter hierarquias de menor número de recursos que as originalmente extraídas. A redução de hierarquias proporciona melhor desempenho computacional em aplicações como dimensões de análise em DW e uso de hierarquias para navegação de dados. Além disso, a omissão de recursos não-relevantes torna hierarquias mais simples e mais exatas para a utilização do usuário final.

Embora a hierarquia adaptada pelo algoritmo *SimpleTailoring* represente todos os dados semanticamente anotados gerados na fase de *Enriquecimento Semântico*, pode-se entender que haja perda de informação ao esconder recursos da hierarquias em nodos "Others". Este algoritmo de adaptação tem o intuito de provocar esta perda de informação, pois considera que os recursos escondidos não sejam relevantes para o consumidor da hierarquia de recursos adaptada. Ainda, caso ocorra uma omissão indesejada, é possível corrigir a hierarquia adaptada durante o Passo 9, adicionando manualmente recursos escondidos.

Para melhor entendimento das consequências da adaptação de hierarquias, experimentos foram realizados utilizando diferentes SMoDs para a fase de *Modelagem de hierarquias* e diferentes limiares para o algoritmo *SimpleTailoring*. Estes experimentos são descritos no capítulo a seguir.

4 AMBIENTE E RESULTADOS EXPERIMENTAIS

Este capítulo apresenta o ambiente experimental adotado para investigar a viabilidade e efeitos do uso do método para adaptação de hierarquias de recursos proposto nesta dissertação. Os resultados experimentais da adaptação de hierarquias de recursos são analisados e discutidos. Os experimentos realizados aplicaram o método proposto a quatro SMoDs distintos, contendo *tweets* semanticamente anotados com recursos LOD sobre lugares de interesse (PoIs) visitados. Por fim, considerações finais discutem o desempenho e viabilidade da implementação e método para grande volumes de dados.

4.1 AMBIENTE EXPERIMENTAL

Os dados alvo de anotações semânticas eleitos para realizar experimentos são dados sobre movimento, devido a trabalhos anteriores do Grupo de Banco de Dados da UFSC (GBD) e de grupos parceiros pelo projeto *Semantic EnrichmEnt of trajectory Knowledge discovery* (SEEK). Como há dificuldade de se obter e criar trajetórias anotadas, os experimentos utilizaram trilhas de usuários na mídia social Twitter, coletados a partir de sua API de dados.

Tweets já foram utilizados em trabalhos anteriores do grupo de pesquisa (MAY; FILETO, 2014; FILETO et al., 2015). Em May e Fileto (2014), anotações semânticas são criadas para anotar *tweets*.

Os recursos de enriquecimento semântico usados nas anotações semânticas de *tweets* descrevem os PoIs visitados pelos usuários e citados no conteúdo textual do *tweet*. Para garantir que o PoIs visitado por um usuário é o mesmo lugar citado no conteúdo de um *tweet*, foram selecionados apenas *tweets* automaticamente gerados pelo sistema Foursquare[1] durante o *check-in* de usuários do sistema em lugares registrados.

Atualmente, há um grande volume de PoIs descritos em KBs. A Tabela 1 mostra o número atual de recursos e triplas de algumas coleções de LODs que descrevem lugares. DBpedia descreve cerca de 735 K PoIs. YAGO2 é uma KB automaticamente extraída da Wikipedia, WordNet e GeoNames, e a acurácia de seus dados foi estimada ser superior a 95% em uma amostra de fatos.

Estas coleções são ricas fontes de dados de pesquisa para a ex-

[1] https://foursquare.com

LOD	DBpedia	GeoNames	LGD	YAGO2
Número de recursos	5.9 M	10 M	3 G	10 M
Número de triplas	6.9 G	150 M	20 G	120 M

Tabela 1 – Dimensão de coleções de LOD

tração de hierarquias de recursos sobre PoIs. Entretanto, é necessaária uma ferramenta capaz de lidar com grande volume de dados de diferentes fontes.

A extração de hierarquias de recursos a partir de conjuntos dados semanticamente anotados e a adaptação de hierarquias foram executadas com auxílio do protótipo ferramental *SeMovDim*, que implementa o método proposto. O ambiente computacional, a ferramenta utilizada para implementar passos do método proposto e a execução da ferramenta são apresentados a seguir.

4.1.1 Ambiente Computacional e Ferramenta SeMovDim

Os experimentos foram executados em uma máquina Intel Core i3-2330M de quatro cores a 2,20 GHz, com 3,8 GB de memória RAM em sistema operacional Ubuntu 15.04 64-bit.

Os conjuntos de dados semanticamente anotados, resultantes da fase de *Enriquecimento Semântico*, são 4 SMoDs armazenados em STDBs gerenciados pelo SGBD PostgreSQL 9.3[2] com extensão PostGIS. Os recursos e conexões extraídos são armazenados no repositório de triplas TDB[3].

O protótipo ferramental *SeMovDim* implementa o método proposto na linguagem Java 1.7. *SeMovDim* foi desenvolvida com auxílio do ambiente de desenvolvimento Eclipse, do *framework Jena*[4], da API *Java Database Conectivity* (JDBC).

4.1.2 Execução do método para adaptação de hierarquias

A fase de *Enriquecimento Semântico* foi realizada em conjunto com colaboradores (MAY; FILETO, 2014; FILETO et al., 2015) e é melhor abordada por estes trabalhos. Os *tweets* foram semanticamente ano-

[2] http://www.postgresql.org/
[3] https://jena.apache.org/documentation/tdb/
[4] http://jena.apache.org/

tados com recursos sobre PoIs visitados descritos nas coleções de LOD DBpedia e LinkedGeoData (LGD). As anotações semânticas foram geradas considerando a adoção de dois critérios: proximidade espacial entre PoIs e *tweets*, e similaridade textual entre nome do lugar e sua menção no conteúdo do *tweet*.

	região	duração	#usuários	#tweets	#PoI
SMoD-1	BR	06-07/14	1.039	1.530	258
SMoD-2	BR	06-07/14	6.343	10.710	1.501
SMoD-3	FLN	10-11/14	468	1.109	110
SMoD-4	BR	06/14-01/15	68.008	327.621	25.079

Tabela 2 – SMoDs que explicitam o lugar visitado no *tweet*

A fase de *Modelagem de Hierarquias* e de *Adaptação de Hierarquias* foram executadas com auxílio do protótipo ferramental *SeMovDim*. Para cada SMoD, foram extraídas duas hierarquias de recursos: uma Hierarquia de recursos sobre objetos, que é composta por relações de ordenamento parcial por conteção espacial (*part of*); e uma hierarquia de recursos sobre conceitos, composta por relações de ordenamento parcial por organização conceitual (*is a*).

Durante o Passo 4, *SeMovDim* extraiu recursos conectados a recursos de enriquecimento semântico das fontes de dados de pesquisa: LGD, DBpedia, GeoNames e *Global Administrative Areas* (GADM).

O Passo 7 foi executado utilizando *SeMovDim* para contar a frequência de uso de cada recursos das hierarquias geradas nas anotações semânticas do SMoD. O Passo 8 aplicou o algoritmo *SimpleTailoring* aplicando diferentes valores de limiar (1, 2, 4, 8, 16, 32, 64, 128, 256, 512 e 1024). O Passo 9 não foi aplicado aos resultados obtidos, mas é possível aplicá-lo com qualquer ferramenta que manipule de forma gráfica ou textual arquivos RDF (*e.g.*, Protege[5]).

A Figura 19 ilustra uma hierarquia de recursos sobre objetos e foi gerada manualmente pelo autor. A Figura 20 ilustra uma hierarquia de recursos sobre conceitos e foi gerada utilizando a ferramenta de visualização do Protege. Ambas ilustrações são extratos de hierarquias geradas durante a execução do método proposto.

As hierarquias adaptadas foram analisadas com o intuito de investigar a viabilidade do método e algoritmo de adaptação propostos. Os resultados experimentais e a análise são descritos a seguir.

[5] http://protege.stanford.edu/

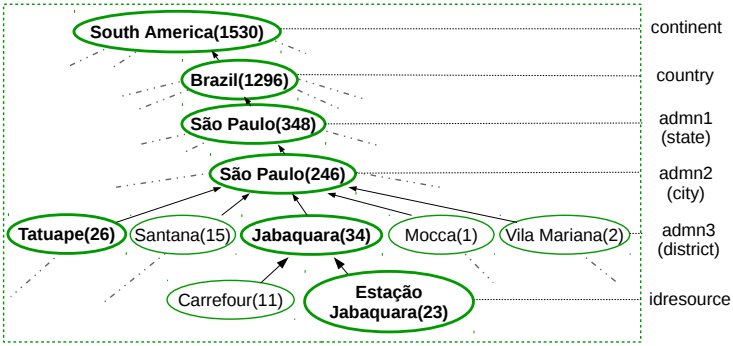


Figura 19 – Extrato da hierarquia de recursos sobre objetos

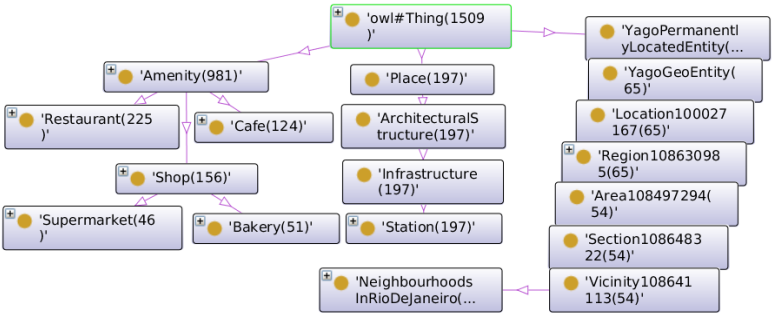


Figura 20 – Extrato da hierarquia de recursos sobre conceitos

4.2 RESULTADOS EXPERIMENTAIS

Experimentos foram realizados comparando as hierarquias de recursos adaptadas para entender as consequências da aplicação do algoritmo *Simple Tailoring* com diferentes limiares inferiores. A análise dos resultados experimentais comparou o número de recursos, variando hierarquias de recursos, limiar de adaptação para o algoritmo *Simple Tailoring* e o nível analisado da hierarquia.

A Figura 21 mostra que o aumento do limiar inferior da frequência de uso causa a redução esperada do número de recursos na hierarquia adaptada.

Além disso, a figura mostra que recursos dos níveis maiores da hierarquia são mais suscetíveis a omissão pelo algoritmo *SimpleTailoring* que recursos de níveis menores. Isto é, a redução do número de

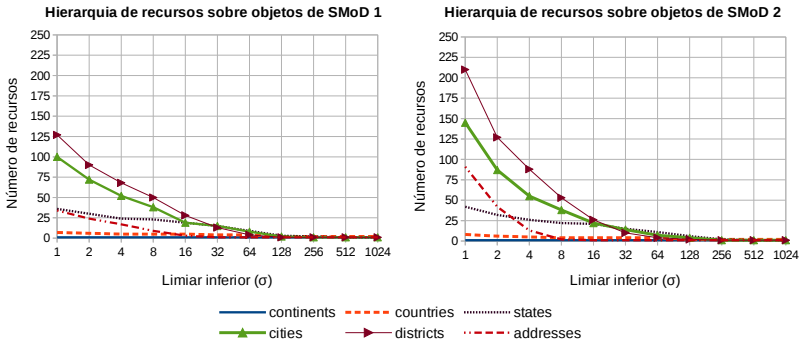


Figura 21 – Número de recursos de cada nível de hierarquias de recursos sobre objetos adaptadas por valores de σ em ordem ascendente

recursos do tipo distrito (nível 4) é maior que a redução dos de tipo cidade (nível 3).

Este comportamento é reflexo das propriedades do número de frequência de uso de recursos, de diferentes níveis de generalização da hierarquia adaptada, em anotações semânticas. A relação direta ou indireta de um recurso de maior nível r' com um dado alvo de anotação t implica na relação indireta de um recurso de nível menor r'' com t, desde que r'' seja ancestral de r' na hierarquia adaptada. Por isso, um recurso possui frequência de uso maior ou igual a frequência de uso de seus recursos descendentes.

4.3 USO DE HIERARQUIAS EM DATA WAREHOUSE

A hierarquia de recursos adaptadas pelo experimento, sobre objetos (Figura 19) e conceitos (Figura 20), podem ser utilizadas como dimensões de análise em DWs. A Figura 22 apresenta um esquema lógico de referência para a construção de MDWs (FILETO et al., 2014). A tabela *FactMSegm* armazena dados análogos aos segmentos de dados brutos sobre movimento. *FactMO* armazena informações a respeito dos objetos móveis. A tabela *Space* liga os fatos de análise as dimensões *SpatialO* e *SpatialC*. A hierarquia de recursos sobre objetos pode ser utilizada para popular a dimensão *SpatialO* e a hierarquia de recursos sobre conceitos para popular a *SpatialC*. Maiores detalhes sobre o esquema são encontrados em Fileto et al. (2014).

Hierarquias de recursos sobre conceitos, quando utilizadas como

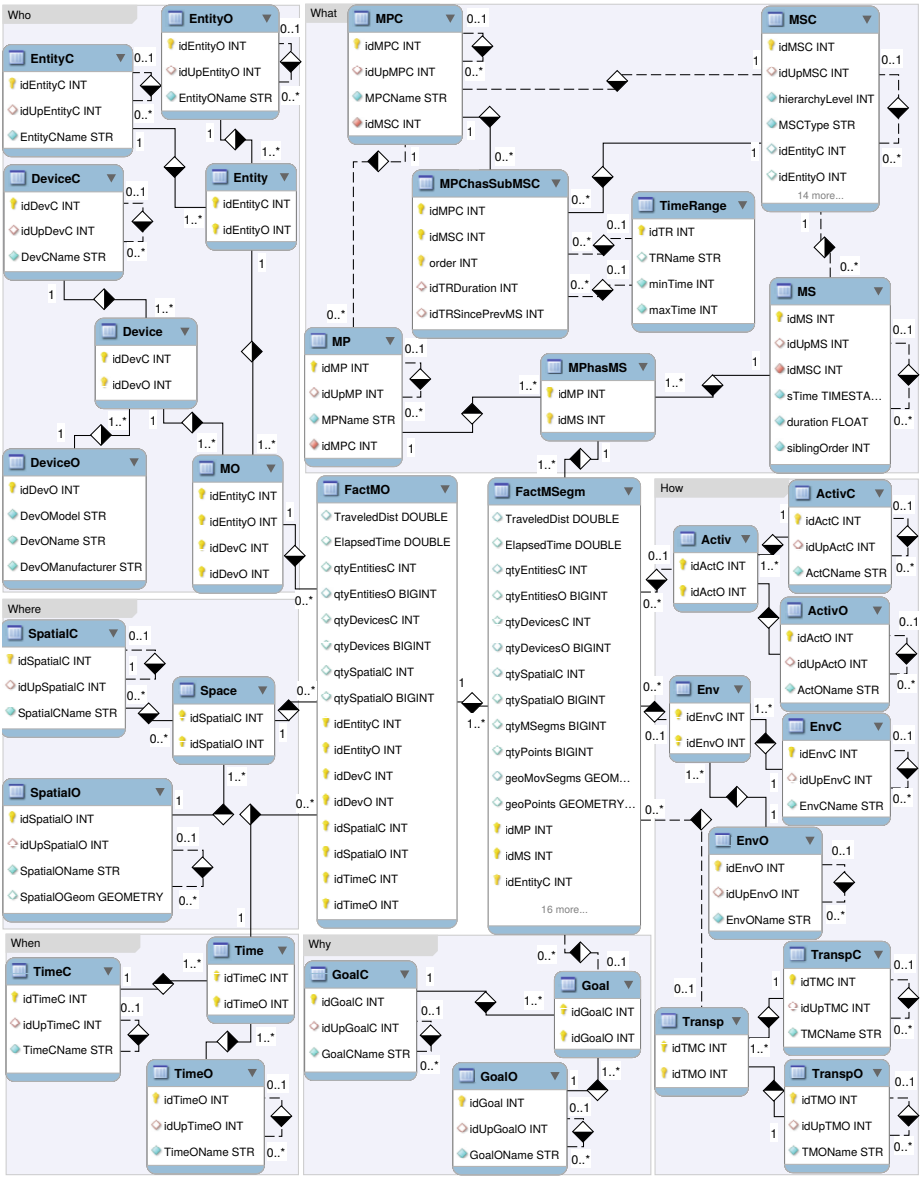


Figura 22 – Esquema lógico de referência para MDW (FILETO et al., 2014)

dimensões de análise, proporcionam novos meios de analisar conjuntos de dados semanticamente anotados. Um MDW construído com base no esquema de referência, populado com os tweets associados a recursos de LOD usados nos experimentos e com as hierarquias de recursos ilustradas pelas (Figuras 19 e 20), pode realizar consultas como:

1. Quais os tipos de lugares mais visitados?

```
SELECT SC.idSpatialC, SC.SpatialCName,
    COUNT(FS.qtyPoints)
FROM FactMSegment FS, FactMO, Space, SpatialC SC
WHERE #<condições_de_junção_natural>
GROUP BY SC.idSpatialC
ORDER BY 3 DESC;
```

2. Quais são os lugares mais visitados do tipo *Restaurant*?

```
SELECT SO.idSpatialO, SO.SpatialOName,
    COUNT(FS.qtyPoints)
FROM FactMSegment FS, FactMO, Space, SpatialO SO,
    SpatialC SC
WHERE #<condições_de_junção_natural>
    AND SC.SpatialCName like "Restaurant"
GROUP BY SC.idSpatialO
ORDER BY 3 DESC;
```

3. Qual o número de tweets de usuários em lugares do tipo *Station*?

```
SELECT COUNT(FS.qtyPoints)
FROM FactMSegment FS, FactMO, Space, SpatialO SO,
    SpatialC SC
WHERE #<condições_de_junção_natural>
    AND SC.SpatialCName like "Station"
GROUP BY SC.idSpatialO
ORDER BY 3 DESC;
```

Resultados da utilização de hierarquias de recursos adaptadas como dimensões de análise em MDW serão apresentadas com maior detalhe na tese (a aparecer em publicação) do colaborador Tommaso Moretto, sob orientação da Prof.ª Dr.ª Alessandra Raffaetà e pela universidade Ca' Foscari de Veneza.

4.4 CONSIDERAÇÕES FINAIS

Esta dissertação não teve o intuito de desenvolver uma ferramenta que implementasse o método proposto de modo eficiente. O protótipo ferramental *SeMovDim* não é capaz de lidar com grande volume de dados, tanto que a extração da hierarquia de recursos sobre conceitos para o SMoD 4 foi abortado devido ao tempo de execução elevado. Entretanto, a demanda contínua e rápida de análise de informação de algumas aplicações, como a análise de mídias sociais (*e.g.*, Twitter, Facebook, Flick), pode exigir uma implementação que explore conceitos de computação paralela e distribuída.

Dentro do escopo desta dissertação, optamos pelo planejamento de estratégias e métodos para a extração e adaptação de hierarquias de recursos. A comparação desta dissertação com trabalhos relacionados é descrita no próximo capítulo.

5 TRABALHOS RELACIONADOS

Este trabalho propôs um método de extração e adaptação de hierarquias de recursos de KBs para gerar dimensões de análise em DW. Este trabalho permite a integração de fonte de dados externa e não convencional (*i.e.* LOD, KBs) e a análise de dados semanticamente anotados.

Os parâmetros selecionados para compará-lo com trabalhos correlatos:

1. Reuso de recursos de LOD e KBs
2. Utilização de hierarquia de recursos
3. Adaptação de hierarquia de recursos
4. Geração de dimensão a partir de hierarquia de recursos
5. Viabilidade de modelar dimensões sobre conceitos e instâncias separadamente

A tabela 3 compara a abordagem proposta e implementada no protótipo *SeMovDim* com abordagens que geram hierarquia de recursos para a construção de DW (DANGER; BERLANGA, 2009; NEUMAYR; SCHREFL; LINNER, 2011; ANDERLIK; NEUMAYR; SCHREFL, 2012; NEBOT; BERLANGA, 2012), que criam meta-modelos para hierarquias de dimensões (NEUMAYR; ANDERLIK; SCHREFL, 2012; GALLINUCCI; GOLFARELLI; RIZZI, 2015), e que de algum modo reduzem o tamanho de dimensões de análise (LIU; IFTIKHAR, 2013).

Com relação ao primeiro critério da Tabela 3 (Reuso de LOD e KB), somente Gallinucci, Golfarelli e Rizzi (2015) não reutilizam recursos de LOD e KBs; entretanto utilizam como dimensões de análise hierarquias de tópicos criadas pelo usuário (analista do DW), representadas com o uso de tecnologias da *Web* Semântica. Neumayr, Schrefl e Linner (2011) reutilizam a ontologia de domínio referenciada pelas anotações semânticas como entrada para auxiliar no processo de interpretação e análise dos dados do DW.

O segundo critério de comparação das abordagens analisa se houve e como foram utilizadas hierarquias de recursos. Danger e Berlanga (2009) propõem uma ferramenta que utiliza conceitos de ontologias (nível intensional) para analisar instâncias destas ontologias (nível extensional). A estrutura de hierarquia de recursos é utilizada para descrever conceitos da ontologia, e cada conceito é enumerado com número

de instâncias. Neumayr, Anderlik e Schrefl (2012) definem e implementam uma estrutura abstrata e semântica de ontologias para representar classes e objetos de interesse expressos pelo usuário (analista do DW) em consultas por meio de expressões númericas (*e.g.*, uma pessoa cujo atributo $idade > 65$ é classificada como instância da classe *idoso*).

Além de Gallinucci, Golfarelli e Rizzi (2015), duas abordagens utilizam hierarquias como dimensões de análise. Anderlik, Neumayr e Schrefl (2012) investigam como anotações semânticas e as ontologias de domínio que elas referenciam podem ser melhor exploradas em análises em DW. Os autores definem um processo de geração de dimensões por meio de hierarquias de conceitos formadas por relações de ordenamento parcial (*subsumption*) e extendem as operações OLAP adaptando-as as dimensões geradas. Já a abordagem de Nebot e Berlanga (2012) definem um método semiautomático que extrai fatos e dimensões de anotações semânticas. Duas medidas são propostas para extrair hierarquias de recursos em um formato que contendo propriedades características de dimensões multidimensionais (*e.g.*, sumarização).

A adaptação de hierarquias (terceiro critério de comparação das abordagens) não é realizada por nenhum dos trabalhos correlatos. Contuto, Liu e Iftikhar (2013) define uma metodologia para particionar dimensões com grande número de atributos e valores de atributos (*big dimensions*).

Os trabalhos de Neumayr, Schrefl e Linner (2011), Anderlik, Neumayr e Schrefl (2012), Nebot e Berlanga (2012), utilizam hierarquias de recursos para gerar dimensões para análise de dados semanticamente anotados (quarto critério de comparação das abordagens). Em Neumayr, Anderlik e Schrefl (2012), o analista cria hierarquias de conceitos a partir de hierarquias de dimensões tradicionais para aprimorar a expressividade de consultas.

Apenas nossa abordagem apresentou um exemplo real contendo dimensões conceituais (*i.e.*, cujos membros são conceitos e não instâncias) e de objetos (cujos membros são instâncias) para um determinado universo de discurso. A separação de conceitos e objetos em dimensões de análise (quinto critério) já foi anteriormente proposta (FILETO et al., 2014) e o estudo de suas implicações é um tema de pesquisa em andamento. Entretanto, a proposta de Nebot e Berlanga (2012) pode ser adaptada para a geração de dimensões de conceitos e de objetos.

O método proposto neste trabalho: i) reutiliza recursos de KBs e LOD para enriquecer dados semanticamente anotados; ii) utiliza hierarquia de recursos na construção do DW; iii) apresenta um algoritmo que permite a adaptação automatizada de hierarquias de recursos, com

base no número de anotações semânticas que referenciam direta ou indiretamente cada recurso; iv) gera dimensões para a análise de dados semanticamente anotados, a partir de hierarquias de recursos; e v) possibilita a separação de dimensões sobre conceitos e instâncias de um mesmo universo de discurso.

Há ainda alguns trabalhos que abordam a publicação de DWs e integração de dimensões, explorando problemas da representação de hierarquias de dimensões. Kämpgen e Harth (2011) analisa estatísticas publicadas como dados conectados e aborda alguns dos problemas envolvendo a integração de esquemas multidimensionais de DW expressados com o vocabulário *RDF Data Cube*[1] no modelo multidimensional. Hierarquias de dimensão descritas utilizando *RDF Data Cube* são representadas por cadeias de propriedades LOD. Etcheverry e Vaisman (2012) propõe um vocabulário capaz de explicitar cada nível e valor de nível de hierarquias de dimensão.

Abelló et al. (2015) apresentam um *survey* sobre OLAP exploratório – processo de analítico adaptado a DW que utilizam dados externos não-convencionais (*e.g.* KBs e LOD). Eles definem critérios de categorização e desafios futuros. De acordo com estes critérios, nosso trabalho (SeMovDim) sugere a construção de um DW semântico-consciente de materialização completa, transformações complexas, atualização periódica, de fontes de dados semiestruturados e extensibilidade de evolução.

Nenhum destes trabalhos fornece meios de adaptar hierarquias de recursos extraídos de uma coleção particular de dados, mais especificamente dados sobre movimento, semanticamente anotados. Este trabalho (SACENTI et al., 2015) é, pelo que sabemos, a primeira proposta que gera dimensões de análise a partir de hierarquias de recursos sobre instâncias e conceitos adaptadas de modo automatizado e extraídas de fontes de dados disponíveis na *Web* (LOD).

[1] https://www.w3.org/TR/vocab-data-cube/

Trabalho	Reuso de LOD e KB	Hierarquias de recursos	Adaptação automatizada	Geração de Dimensão	Separação entre conceitos e instâncias
DANGER; BERLANGA, 2009	√	√		√	
NEUMAYR; SCHREFL; LINNER, 2011	√			√	
ANDERLIK; NEUMAYR; SCHREFL, 2012	√	√		√	
NEBOT; BERLANGA, 2012	√	√		√	√
NEUMAYR; ANDERLIK; SCHREFL, 2012	√	√		*	
LIU; IFTIKHAR, 2013	√		*		
GALLINUCCI; GOLFARELLI; RIZZI, 2015				√	
SeMovDim	√	√	√	√	√

Tabela 3 – Tabela comparativa de trabalhos correlatos

6 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresenta avanços na construção de DW para a análise de dados de movimento semânticamente anotados (SMoD). Ele propõe um método para a geração automatizada de dimensões de análise a partir de hierarquias dos recursos (sobre instâncias e conceitos) adaptadas, extraídas de diversas fontes de dados e relacionadas com dados alvo das anotações semânticas analisadas. As principais contribuições são: (i) um método para a extração de hierarquias de recursos de LOD definidas por relações de ordenação parcial (*e.g.*, *is a* e *part of*); (ii) um algoritmo para a adaptação de hierarquias de recursos extraídas; e (iii) a análise dos efeitos da adaptação de hierarquias de recursos em um estudo de caso que gera hierarquias de recursos (sobre objetos e conceitos) a respeito de lugares de interesse (*Place of Interest* - PoI) visitados por usuários do Twitter durante a publicação de *tweets*.

O método proposto neste trabalho: i) reutiliza recursos de KBs e LOD para enriquecer dados semanticamente anotados; ii) extrai hierarquia de recursos relacionados aos recursos de enriquecimento semântico; iii) apresenta um algoritmo que permite a adaptação automatizada de hierarquias de recursos, com base no número de anotações semânticas que referenciam direta ou indiretamente cada recurso; iv) gera dimensões para a análise de dados semanticamente anotados, a partir de hierarquias de recursos; e v) possibilita a separação de dimensões sobre conceitos e instâncias de um mesmo universo de discurso.

Experimentos aplicaram o método proposto a *tweets* semanticamente anotados com recursos de LOD do DBpedia e do LinkedGeoData. Os resultados mostraram que a adaptação de hierarquias produz considerável redução no número de recursos mesmo para limiares baixos de frequência de uso (número de segmentos de movimento relacionados direta ou indiretamente a um recurso). Além disso, experimentos mostraram que recursos de níveis maiores são mais suscetíveis a omissão pelo algoritmo de adaptação proposto.

Este trabalho é, pelo que sabemos, a primeira proposta que gera dimensões de análise a partir de hierarquias de recursos (sobre instâncias e conceitos) adaptadas e extraídas de fontes de dados disponíveis na *Web* (LOD). Resultados parciais desse trabalho foram publicados em um artigo completo na conferência internacional *Big Data Analytics and Knowledge Discovery* (DaWaK) (SACENTI et al., 2015).

A experiência demonstrou-nos que geralmente reutilizar a informação sobre os recursos de enriquecimento semântico para extrair

hierarquias de recursos facilita a construção de dimensões de análise e do DW. Por exemplo, embora a dimensão *Spatial Object Dim* pode ser gerada por meio da aplicação de funções espaciais de contenção em dados geográficos, é também possível extrair uma hierarquia de recursos sobre objetos, de diversas coleções de LOD que descrevem lugares, apenas identificando quais as relações de ordenamento parcial que expressam contenção espacial. Entretanto, algumas vezes é necessário complementar a informação disponível em LOD e KBs por outros meios (*e.g.*, RDB, STDB).

Os principais pontos que não se conseguiu abordar adequadamente no âmbito desta dissertação e ficam para trabalhos futuros são:

1. Estudos teóricos mais aprofundados sobre as implicações do uso de instâncias, conceitos e relações semãnticas em dimensões de DWs;

2. Comparação do método proposto para a geração de dimensões via adaptação de hierarquias oriundas de ontologias e coleções de LOD com outras propostas da literatura;

3. Desenvolvimento de um MDW com dimensões de análise geradas a partir de hierarquias adaptadas;

4. Realização de experimentos com outras bases de dados semanticamente enriquecidos.

5. Investigação de outras bases de conhecimento que contribuam para enriquecer SMoDs anotados com PoIs visitados;

6. Desenvolvimento de uma ferramenta que implemente o método proposto de modo eficiente, explorando conceitos da computação paralela e distribuída, e proporcionando atualização de dimensões de DWs em tempo real;

7. Investigação mais aprofundada dos efeitos da adaptação de hierarquias de recursos, tanto no desempenho computacional quanto na sua facilidade de uso pelo usuário;

8. Investigação de outras aplicações para hierarquias de recursos.

REFERÊNCIAS

ABELLÓ, A. et al. Using semantic web technologies for exploratory OLAP: A survey. **IEEE Trans. Knowl. Data Eng.**, v. 27, n. 2, p. 571–588, 2015. Disponível em: <http://dx.doi.org/10.1109/TKDE.2014.2330822>.

ANDERLIK, S.; NEUMAYR, B.; SCHREFL, M. Using domain ontologies as semantic dimensions in data warehouses. In: ATZENI, P.; CHEUNG, D. W.; RAM, S. (Ed.). **ER**. Springer, 2012. (Lecture Notes in Computer Science, v. 7532), p. 88–101. ISBN 978-3-642-34001-7. Disponível em: <http://dblp.uni-trier.de/db/conf/er/er2012.html#AnderlikNS12>.

ANGLES, R.; GUTIERREZ, C. **The expressive power of SPARQL**. [S.l.]: Springer, 2008.

BERNERS-LEE, T. **Uniform Resource Identifier (URI): Generic Syntax**. 2005. Disponível em: <http://tools.ietf.org/html/rfc3986>.

BERNERS-LEE, T. **Linked Data - Design Issues**. 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: Scientific American. **Scientific American**, v. 284, n. 5, maio 2001. Disponível em: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&pageNumber=1&catID=2>.

BOGORNY, V. et al. CONSTAnT - A Conceptual Data Model for Semantic Trajectories of Moving Objects. **T. GIS**, v. 18, n. 1, p. 66–88, 2014.

CABIBBO, L.; TORLONE, R. Querying multidimensional databases. In: **Database programming languages**. [S.l.]: Springer, 1998. p. 319–335.

DANGER, R.; BERLANGA, R. A semantic web approach for ontological instances analysis. In: FILIPE, J. et al. (Ed.). **Software and Data Technologies**. Springer Berlin Heidelberg, 2009, (Communications in Computer and Information Science, v. 22). p.

269–282. ISBN 978-3-540-88654-9. Disponível em: <http://dx.doi.org/10.1007/978-3-540-88655-6_20>.

DELCAMBRE, L. M. L.; MAIER, D. Models for superimposed information. In: **Proceedings of the Workshops on Evolution and Change in Data Management, Reverse Engineering in Information Systems, and the World Wide Web and Conceptual Modeling**. London, UK, UK: Springer-Verlag, 1999. (ER '99), p. 264–280. ISBN 3-540-66653-2. Disponível em: <http://dl.acm.org/citation.cfm?id=647523.728336>.

ETCHEVERRY, L.; VAISMAN, A. A. Enhancing olap analysis with web cubes. In: **Proceedings of the 9th International Conference on The Semantic Web: Research and Applications**. Berlin, Heidelberg: Springer-Verlag, 2012. (ESWC'12), p. 469–483. ISBN 978-3-642-30283-1. Disponível em: <http://dx.doi.org/10.1007/978-3-642-30284-8_38>.

ETCHEVERRY, L.; VAISMAN, A. A.; ZIMÁNYI, E. Modeling and querying data warehouses on the semantic web using QB4OLAP. In: **DaWaK**. [S.l.: s.n.], 2014. (LNCS, v. 8646), p. 45–56.

FILETO, R. et al. Semantic enrichment and analysis of movement data: probably it is just starting! **SIGSPATIAL Special**, v. 7, n. 1, p. 11–18, 2015. Disponível em: <http://doi.acm.org/10.1145/2782759.2782763>.

FILETO, R. et al. Baquara: A Holistic Ontological Framework for Movement Analysis Using Linked Data. In: **ER**. [S.l.: s.n.], 2013. p. 342–355.

FILETO, R. et al. The baquara^2 knowledge-based framework for semantic enrichment and analysis of movement data. **Data Knowl. Eng.**, v. 98, p. 104–122, 2015. Disponível em: <http://dx.doi.org/10.1016/j.datak.2015.07.010>.

FILETO, R. et al. A semantic model for movement data warehouses. In: **DOLAP 2014**. [S.l.: s.n.], 2014. p. 47–56.

GALLINUCCI, E.; GOLFARELLI, M.; RIZZI, S. Meta-stars: Dynamic, schemaless, and semantically-rich topic hierarchies in social BI. In: **18th Intl. Conf. on Extending Database Technology, EDBT 2015, Brussels**. [s.n.], 2015. p. 529–532. Disponível em: <http://dx.doi.org/10.5441/002/edbt.2015.50>.

GOLFARELLI, M.; MAIO, D.; RIZZI, S. The dimensional fact model: A conceptual model for data warehouses. **International Journal of Cooperative Information Systems**, v. 7, p. 215–247, 1998.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. **Int. J. Hum.-Comput. Stud.**, Academic Press, Inc., Duluth, MN, USA, v. 43, n. 5-6, p. 907–928, dez. 1995. ISSN 1071-5819. Disponível em: <http://dx.doi.org/10.1006/ijhc.1995.1081>.

GUARINO, N. Formal ontology and information systems. In: . [S.l.]: IOS Press, 1998. p. 3–15.

HONG, L. et al. Discovering geographical topics in the twitter stream. In: **Proceedings of the 21st International Conference on World Wide Web**. New York, NY, USA: ACM, 2012. (WWW '12), p. 769–778. ISBN 978-1-4503-1229-5. Disponível em: <http://doi.acm.org/10.1145/2187836.2187940>.

INMON, W. H.; STRAUSS, D.; NEUSHLOSS, G. **DW 2.0: The Architecture for the Next Generation of Data Warehousing**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008. ISBN 0123743192, 9780123743190.

KäMPGEN, B.; HARTH, A. Transforming statistical linked data for use in olap systems. In: **Proceedings of the 7th International Conference on Semantic Systems**. New York, NY, USA: ACM, 2011. (I-Semantics '11), p. 33–40. ISBN 978-1-4503-0621-8. Disponível em: <http://doi.acm.org/10.1145/2063518.2063523>.

KIMBALL, R. **The Data Warehouse Toolkit**. [S.l.]: J. Wiley and Sons, 1996.

LIU, X.; IFTIKHAR, N. Ontology-based big dimension modeling in data warehouse schema design. In: ABRAMOWICZ, W. (Ed.). **Business Information Systems**. Springer Berlin Heidelberg, 2013, (Lecture Notes in Business Information Processing, v. 157). p. 75–87. ISBN 978-3-642-38365-6. Disponível em: <http://dx.doi.org/10.1007/978-3-642-38366-3_7>.

Rdf primer. Februar 2004. Stand: 15.4.2009. Disponível em: <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>.

MAY, C.; FILETO, R. Connecting Textually Annotated Movement Data with Linked Data. In: **IX Regional School on Databases**.

São Francisco do Sul, SC, Brazil (in Portuguese): SBC, 2014. (ERBD).

NEBOT, V.; BERLANGA, R. Building data warehouses with semantic web data. **Decis. Support Syst.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 52, n. 4, p. 853–868, mar. 2012. ISSN 0167-9236. Disponível em: <http://dx.doi.org/10.1016/j.dss.2011.11.009>.

NEUMAYR, B.; ANDERLIK, S.; SCHREFL, M. Towards ontology-based olap: Datalog-based reasoning over multidimensional ontologies. In: **Proceedings of the Fifteenth International Workshop on Data Warehousing and OLAP**. New York, NY, USA: ACM, 2012. (DOLAP '12), p. 41–48. ISBN 978-1-4503-1721-4. Disponível em: <http://doi.acm.org/10.1145/2390045.2390053>.

NEUMAYR, B.; SCHREFL, M.; LINNER, K. Semantic cockpit: An ontology-driven, interactive business intelligence tool for comparative data analysis. In: TROYER, O. D. et al. (Ed.). **Advances in Conceptual Modeling. Recent Developments and New Directions**. Springer Berlin Heidelberg, 2011, (Lecture Notes in Computer Science, v. 6999). p. 55–64. ISBN 978-3-642-24573-2. Disponível em: <http://dx.doi.org/10.1007/978-3-642-24574-9_9>.

OREN, E. et al. **What are Semantic Annotations?** 2006.

PARDILLO, J.; MAZÓN, J.-N. Using ontologies for the design of data warehouses. **arXiv preprint arXiv:1106.0304**, 2011.

PARENT, C. et al. Semantic trajectories modeling and analysis. **ACM Comput. Surv.**, v. 45, n. 4, 2013. Article 42.

PELEKIS, N.; THEODORIDIS, Y. **Mobility Data Management and Exploration**. [S.l.]: Springer, 2014. 1-298 p. ISBN 978-1-4939-0391-7, 978-1-4939-0392-4.

PERRY, M.; SHETH, A.; JAIN, P. **SPARQL–ST: Extending SPARQL to Support Spatiotemporal Queries, Kno. e**. [S.l.], 2008.

RIGAUX, P.; SCHOLL, M.; VOISARD, A. **Introduction to Spatial Databases: Applications to GIS**. [S.l.]: Morgan Kaufmann, 2000. ISBN 1-55860-689-0.

RINZIVILLO, S. et al. Where Have You Been Today? Annotating Trajectories with DayTag. In: **SSTD**. [S.l.]: Springer, 2013. (LNCS, v. 8098), p. 467–471.

SACENTI, J. A. P. et al. Automatically tailoring semantics-enabled dimensions for movement data warehouses. In: **DAWAK 2015**. [S.l.: s.n.], 2015.

YAN, Z. et al. Semantic trajectories: Mobility data computation and annotation. **ACM TIST**, v. 4, n. 3, 2013.